


## ORIGINAL RESEARCH

# Object detection in smart indoor shopping using an enhanced YOLOv8n algorithm

Yawen Zhao<sup>1</sup> | Defu Yang<sup>1</sup> | Sheng Cao<sup>1</sup> | Bingyu Cai<sup>1,2</sup> | Maryamah Maryamah<sup>3</sup> | Mahmud Iwan Solihin<sup>1</sup> 

<sup>1</sup>Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur, Malaysia

<sup>2</sup>School of Advanced Manufacturing, Shantou Polytechnic, Shantou, China

<sup>3</sup>Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya, Indonesia

## Correspondence

Mahmud Iwan Solihin, Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur, Malaysia.  
Email: mahmudis@ucsiuniversity.edu.my

## Abstract

This paper introduces an enhanced object detection algorithm tailored for indoor shopping applications, a critical component of smart cities and smart shopping ecosystems. The proposed method builds on the YOLOv8n algorithm by integrating a ParNetAttention module into the backbone's C2f module, creating the novel C2f-ParNet structure. This innovation enhances feature extraction, crucial for detecting intricate details in complex indoor environments. Additionally, the channel-wise attention-recurrent feature extraction (CARAFE) module is incorporated into the neck network, improving target feature fusion and focus on objects of interest, thereby boosting detection accuracy. To optimize training efficiency, the model employs the Wise Intersection over Union (WIoUv3) as its regression loss function, accelerating data convergence and improving performance. Experimental results demonstrate the enhanced YOLOv8n achieves a mean average precision (mAP) at 50% threshold (mAP@50) of 61.2%, a 1.2 percentage point improvement over the baseline. The fully optimized algorithm achieves an mAP@50 of 65.9% and an F1 score of 63.5%, outperforming both the original YOLOv8n and existing algorithms. Furthermore, with a frame rate of 106.5 FPS and computational complexity of just 12.9 GFLOPs (Giga Floating-Point Operations per Second), this approach balances high performance with lightweight efficiency, making it ideal for real-time applications in smart retail environments.

## 1 | INTRODUCTION

Intelligent shopping assistants, empowered by robotics and deep learning, promise transformative solutions to enhance the shopping experience [1], particularly for individuals with visual or mobility impairments. These innovations are becoming increasingly pertinent as artificial intelligence continues

to evolve and integrate into various applications [2–4]. The core functionality of these robots hinges on precise object detection—an integral component of machine vision and a fundamental aspect of robotic navigation and interaction [5]. Object detection, a critical subset of machine vision, is essential for the seamless integration of deep learning technologies with robotics, utilizing RGB depth cameras to enhance operational efficiency [6]. In recent advancements, Jiang et al. [7] have designed a supermarket shopping robot that employs a STM32F407 microcontroller as its core processing unit. This design leverages a modified AlexNet architecture for robust object feature extraction, aiming to expedite the recognition process of various items. Despite its innovative approach, the system faced challenges such as false positives and missed detections during real-world trials. Additionally, issues with camera image acquisition, primarily blurring, were observed, which significantly hampered the accuracy of object recognition.

**Abbreviations:** CARAFE, Channel-wise attention-recurrent feature extraction; CIoU, Complete intersection over union; CNN, Convolutional neural networks; DIoU, Distance-IoU; EIoUv1, Efficient-IoUv1; FM, Focusing mechanism; FPN, Feature pyramid network; FPS, Frames per second; GIoU, Generalized-IoU; Grad CAM, Gradient-weighted class activation mapping; IoU, Intersection over union; mAP, Mean average precision; PAFPN, Path aggregation feature pyramid network; PAN, Path aggregation network; ROI, Region of interest; RPN, Region proposal network; SPP, Spatial pyramid pooling; SPPF, Spatial pyramid pooling function; SSD, Single shot multibox detector; SSE, Skip squeeze and excitation; S-SSD, ShuffleNetSSD; WIoU, Wise IoU; WIoUv3, Wise intersection over union; YOLO, You Only Look Once; YOLOv8-CPN-CW, “CPN” denotes the ParNet\_C2f module, “C” represents the CARAFE module, and “W” represents the WIoUv3 module.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *IET Image Processing* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

In another study, Zhang et al. [8] presented an image recognition algorithm for shopping robots based on deep learning convolutional neural networks (CNN), overcoming issues related to low recognition accuracy and slow recognition speed. Nevertheless, the study lacks comprehensive testing in diverse scenarios. In another approach, Aydin et al. [9] introduced an object detection system for indoor robots utilizing a single low-cost camera, demonstrating efficiency through various tests in different environments. For faster object detection, Jiang L et al. [10] introduced a novel lightweight object detection framework designated as ShuffleNetSSD (S-SSD). This framework innovatively substitutes the VGG-16 network with the streamlined SSD algorithm, thereby curtailing computational demands while boosting the efficacy of real-time object detection tailored for indoor robotic applications. Despite these advancements, the model confronts persistent challenges inherent in dynamic environments. These include the detection of multiple targets that exhibit similar colouration, complications arising from target occlusion, and errors induced by robot motion. These factors collectively contribute to occasional missed detections and false positives, which are critical reliability concerns during operational deployment. The ongoing development and refinement of detection systems like S-SSD are crucial, particularly for applications such as shopping assistance robots. These systems hold profound implications for enhancing service safety and accessibility for vulnerable populations including the elderly and individuals with disabilities. As such, the domain of object detection not only demands continuous technological innovations but also a deeper understanding of the interaction dynamics between autonomous systems and human environments.

The integration of shopping robots with deep learning technology has significantly advanced intelligent detection capabilities. Object detection, a crucial aspect of this technology, is broadly categorized into two-stage detection and one-stage detection methodologies. Two-stage detection algorithms, including R-CNN, Fast R-CNN, Faster R-CNN, U-net, etc., are notable for their high accuracy but come with the trade-off of increased recognition time [1].

Enhanced detection accuracy typically involves a trade-off with reduced detection speeds. However, single-stage object detection algorithms challenge this norm by classifying and regressing bounding boxes directly from images, thereby conserving computational resources. Prominent among these are the single shot multibox detector (SSD), You Only Look Once (YOLO) series, and others [11–13].

The SSD algorithm offers an end-to-end solution that performs object category recognition and bounding box detection concurrently. Despite its efficiency, SSD encounters difficulties in adapting to variations in object size and shape, which can yield less than optimal detection in specific scenarios. The YOLO series has seen several iterations, including YOLOv3 [14], YOLOv4 [15], YOLOv5 [16], YOLOv7 [17], and YOLOv8 [18], each enhancing the framework's capabilities. YOLOv3 introduces a feature pyramid network (FPN) to harness multi-scale target information, although it still struggles with localization accuracy and the challenge of detecting multi-

ple or occluded targets. YOLOv4 advances the architecture with CSPDarknet53, integrating spatial pyramid pooling (SPP) and feature fusion networks, which enhance model compression and reduce computational load. However, these improvements demand high computational resources and extensive training periods. YOLOv5 further develops the backbone network with BottleneckCSP and focus modules, incorporating SPP from YOLOv4, and a sophisticated head network that offers enriched feature information. Despite these advancements, YOLOv5's reliance on high-quality datasets can restrict its generalizability across various scenes or tasks.

YOLOv7 utilizes a traditional path aggregation feature pyramid network (PAFPN) structure in the neck module to improve feature fusion from different levels. However, challenges still exist in object detection accuracy. YOLOv8 takes a different approach by moving away from the anchor-based method and embracing the anchor-free concept. With a lightweight network and optimized computational capacity, YOLOv8 achieves higher detection accuracy, faster inference speed, and improved generalization. As a result, compared to two-stage algorithms, YOLOv8 excels in faster detection speeds with fewer parameter calculations, and compared to the SSD algorithm, it demonstrates enhanced object detection capabilities while maintaining a lightweight network architecture.

YOLOv8 represents a single-stage object detection algorithm, distinguished into YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x categories based on the network's depth. Within these classifications, YOLOv8n stands out for its lower computational complexity, fewer parameters, and a more compact model size compared to its counterparts. In the specific context of addressing challenges in object detection for shopping robot operations, issues such as motion-induced blurriness, occlusions of multiple objects, and colour similarity between objects pose significant hurdles. This paper presents an improved YOLOv8n object detection algorithm for indoor shopping robot applications. The key contributions of this paper are outlined as follows:

1. This study introduces the C2f-ParNet module, a novel integration of the ParNetAttention mechanism into the C2f module of the YOLOv8n's backbone network. This hybrid module is designed to enhance the network's capability for nonlinear feature modelling, which significantly improves the extraction and discrimination of complex features in object detection tasks.
2. Then it proposes the replacement of the traditional Upsample module with the improved channel-wise attention-recurrent feature extraction (CARAFE) module within the neck architecture of YOLOv8n. CARAFE employs a sophisticated pixel interpolation strategy, which minimizes information loss during the feature upsampling process. This method not only preserves finer detail but also enhances the perceptual quality of the model's output.
3. The WIoUv3 (Wise Intersection over Union v3) into YOLOv8n marks a significant advancement in the accuracy of object detection. WIoUv3 offers an improved metric for assessing the alignment between predicted and actual

**TABLE 1** Summary of existing object detection benchmarks in retail stores.

Datasets	Number of images	Category	Number of instances	Size	Task	Year
D2S [19]	21,000	60	72,447	1920 × 1440	M	2018
RPC [20]	83,739	200	421,674	1800 × 1800	M	2019
TGFS [21]	38,027	24	38,027	480 × 640	M	2019
Sku110k [22]	11,762	1	1,733,711	1920 × 2560	S	2019
Locount [23]	50,394	140	1,905,317	1920 × 1080	M	2020

bounding boxes. This enhancement is particularly beneficial in complex detection environments where multiple overlapping objects or parts must be accurately identified and delineated.

The paper is organized as follows: Section 1 provides an overview of the research background, outlines existing challenges, and introduces the proposed object detection algorithm specifically designed for shopping robots. Section 2 engages in a detailed discussion of relevant datasets, experimental setups and evaluation metrics, and offers a comprehensive description of the proposed methodology. Section 3 presents the experimental results of the enhanced algorithm, including performance comparisons across various loss functions, ablation studies, and evaluations against related algorithms. Following this, the paper explores visual scene analysis, addressing heatmap analysis and examining real-world scenarios. Finally, Section 4 concludes with the key research results and implications.

## 2 | MATERIALS AND METHODS

### 2.1 | Supermarket object detection dataset

For comprehensive supermarket object detection, it is crucial to select a dataset that exhibits strong generalization and applicability across various scenarios. The dataset's types, quantities, and instances must align with the requirements for subsequent training, testing, and validation stages. To address this, Table 1 shows the insights of datasets from existing literature.

In Table 1, “S” indicates a single-class object detection task, while “M” represents a multi-class object detection task. The Locount dataset meets the specified criteria, featuring an extensive collection of over 140 categories and 1,905,317 instances, which exceeds the coverage of other datasets. Furthermore, the classification levels of the Locount dataset fulfil the requirements for multi-object detection tasks. Consequently, this paper selects the Locount dataset as the preferred choice for supermarket object detection. Examples of images from this dataset are shown in Figure 1.

The Locount dataset comprises images spanning 140 distinct categories, including beverages, shampoo, food, daily necessities, clothing, and electrical appliances. In Figure 2, the dataset's manually labelled information is illustrated across several panels. Figure 2a displays the distribution of object quantities per category, where the predominant category, marked in yellow, encompasses over 4000 instances. Figure 2b details the

dimensions of object bounding boxes, indicating a central concentration around a singular point. Figure 2c elucidates the spatial distribution of bounding box centre points, which predominantly cluster in the depicted black area, primarily towards the dataset's upper and lower extremities. Figure 2d offers a scatter plot of the bounding boxes' widths and heights, with darker hues in the lower-left quadrant suggesting a preponderance of small to medium-sized objects that are prone to occlusion.

To enhance dataset utility under varied conditions, 5000 images exhibiting features such as occlusion, similar coloration, and blurriness were randomly selected from a total of 50,394 images. These images span 109 categories. The selected subset was then arbitrarily divided into training, validation, and testing sets, adhering to a 7:2:1 ratio, to facilitate robust experimental evaluation. Image augmentation processes, including cropping, noise addition, and other transformations were employed on these 5000 images to foster clarity, augment the dataset's robustness, and improve generalization capabilities across diverse visual recognition tasks. This processing yielded a total of 12,560 images, with 8792 designated for training, 2512 for testing, and 1256 for validation purposes.

### 2.2 | Architecture of YOLOv8n

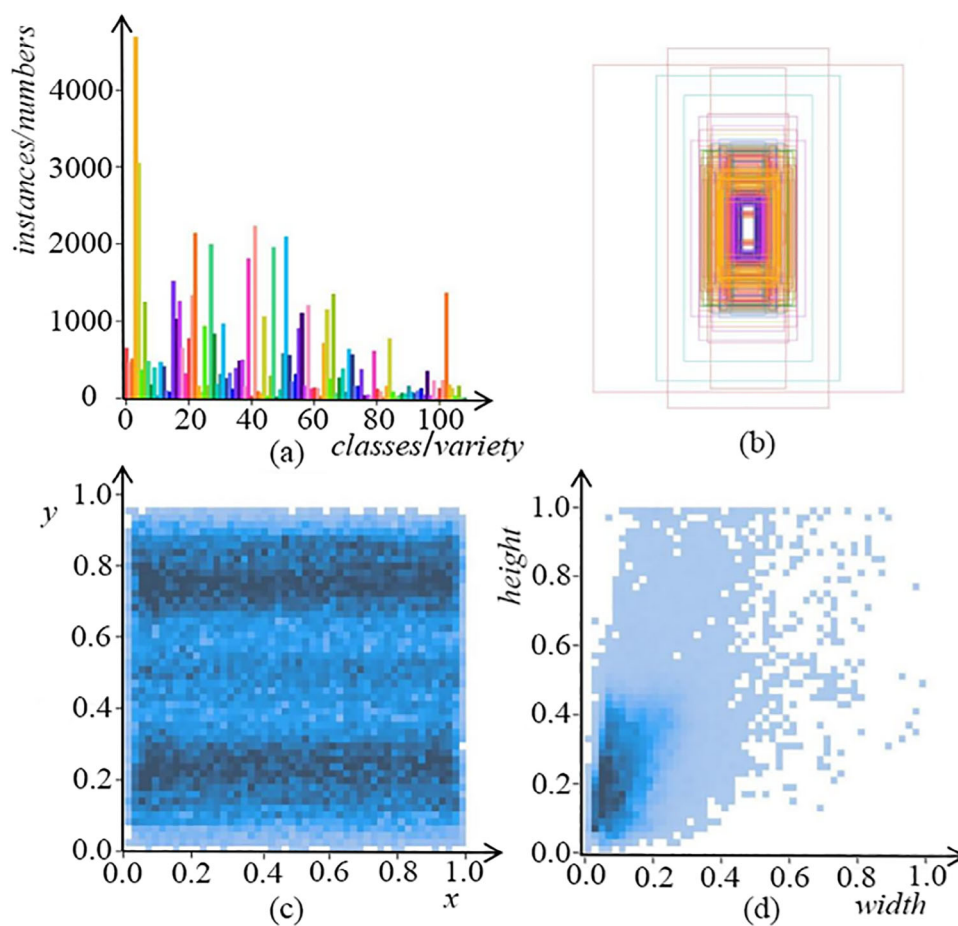
The YOLOv8n model is comprised of input, backbone, neck, head, and output components, with the overall structure depicted in Figure 3. The input receives two-dimensional images of size 1920 × 1080 from the Locount dataset. The backbone network is mainly composed of the C2f structure, Conv module, and SPPF module. The Conv structure primarily extracts features of interest from the input image, using convolutional templates for convolution, batch normalization (BN), and SiLU activation function operations, resulting in images of sizes 640 × 640, 640 × 320, and 320 × 320.

The C2f architecture within the backbone network integrates residual connections to augment the feature fusion capacity of convolutional neural networks (CNNs), thereby facilitating a richer gradient flow. Concurrently, the SPPF module transforms the resultant feature map post-convolution into fixed-size feature vectors, enhancing spatial invariance. The neck network employs a ResNet-based structure, which is instrumental in amalgamating multi-scale features, thereby constructing a robust FPN as described in [24]. This integration of FPN with the path aggregation network (PAN) [25] synergistically merges top-down and bottom-up pathways within the network, significantly bolstering the overall detection capabilities.





**FIGURE 1** Examples of images from the Locount dataset: (a) and (d) are images with motion blur, (b) and (e) are images with the same colour but different categories, (c) and (f) are images with multiple object occlusions.



**FIGURE 2** Information of manually labelled objects in the Locount dataset.

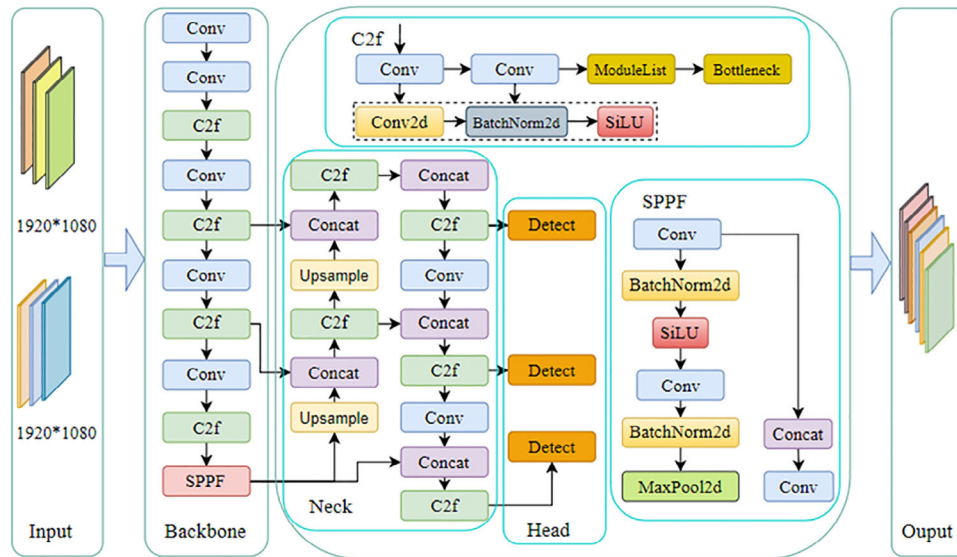


FIGURE 3 The YOLOv8n network structure.

The C2f structure in the neck network does not contain residual structures, mainly aiming to reduce computational complexity and improve network efficiency. The head network facilitates the delineation of differing target objects by correlating category and positional information across various feature map dimensions. This architectural stratification ensures a comprehensive processing pipeline that dynamically adjusts to the object scale variations inherent in complex visual scenes.

While YOLOv8n offers a lightweight architecture aimed at enhancing detection accuracy, it experiences a decrement in processing speed, especially when applied to complex target configurations exemplified by the Locount dataset. This dataset is characterized by challenges such as numerous similarly coloured objects, significant occlusion of small to medium-sized targets, and swift variations in scenes caused by robot motion. Consequently, maintaining robust detection accuracy within such dynamic contexts is imperative for the efficacy of the model.

To tackle the challenges of false positives and missed detections in object detection for shopping robots, the author suggests the implementation of an enhanced YOLOv8-CPN-CW algorithm. This algorithm incorporates the C2f-ParNet module as an improved backbone network denoted by 'CPN,' utilizes the CARAFE module for the neck network denoted by 'C' for enhanced performance, and integrates the WIoUv3 loss function denoted by 'W' for improved accuracy.

## 2.3 | Improvement of YOLOv8n network structure

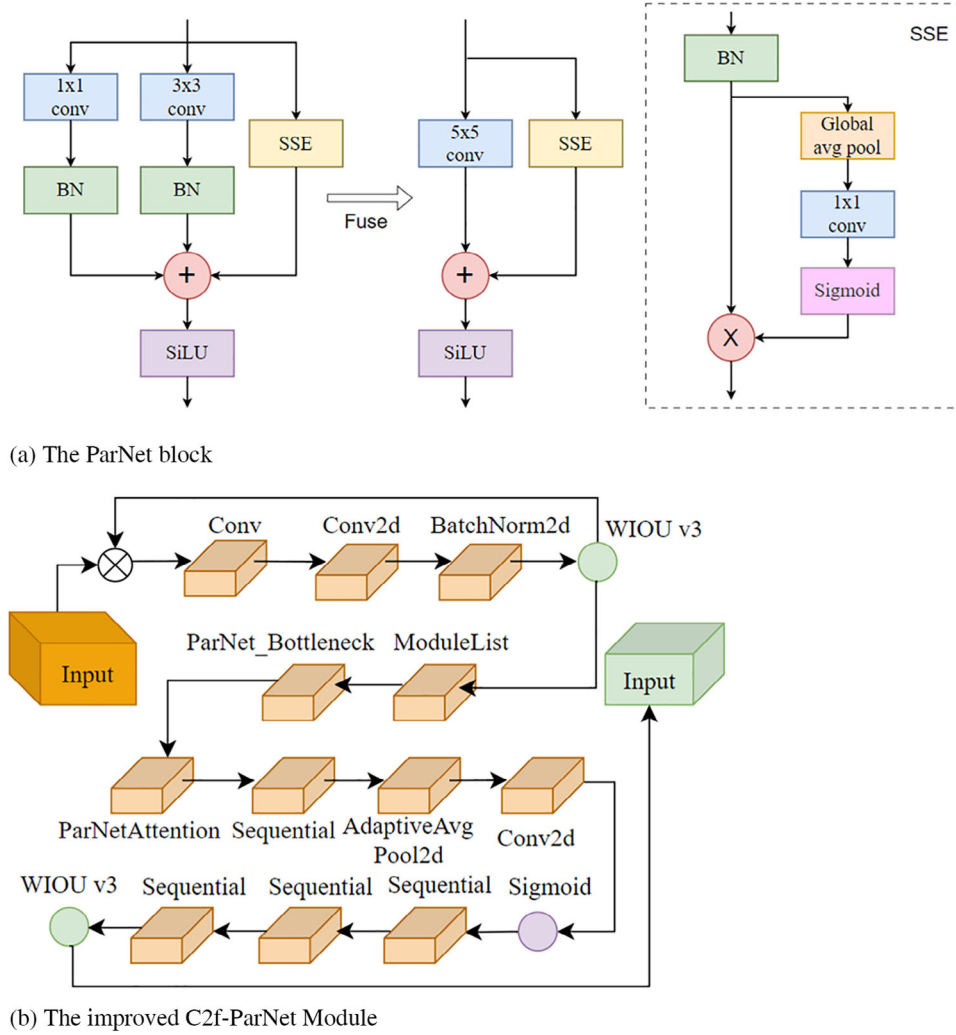
### 2.3.1 | C2f-ParNet module

In order to address the challenges posed by blurred backgrounds and object occlusions common in supermarket envi-

ronments, this study develops a novel approach by integrating the ParNetAttention [26] mechanism into the established C2f structure of the YOLOv8n backbone network, resulting in the innovative C2f-ParNet module (refer to Figure 4b). This enhancement empowers the model to selectively focus on regions of interest while effectively minimizing the impact of noise and irrelevant details, thereby refining the accuracy of object detection.

ParNetAttention introduces a parallel feature extraction module, as depicted in Figure 4a, which comprise  $1 \times 1$  convolution,  $3 \times 3$  convolution, and a skip-squeeze-and-excitation (SSE) mechanism. Upon completion of training, the  $1 \times 1$  and  $3 \times 3$  convolutions can be fused to enhance inference speed. The SSE branch effectively expands the receptive field without impacting the depth of the feature maps. As illustrated in Figure 4b, the bottleneck structure of the ParNet module is designed for efficiency and effectiveness, featuring two convolutional layers and a single ParNetAttention module. Each convolutional layer is meticulously configured with parameters such as the number of groups 'g' in the convolutional kernels and the kernel size 'k', with default settings typically at 1 and  $3 \times 3$ , respectively. During the forward propagation, the initial convolutional layer extracts primary features which are then intricately processed by the subsequent convolutional layer and the ParNetAttention module, culminating in a composite output of feature values through additive operations.

C2f-ParNet is a refined CSP Bottleneck architecture, consisting of two convolutional layers and several ParNet bottleneck modules. The initial convolutional layer captures the relevant features, which are then split into two segments using the split() function. One segment is fed into the ParNet bottleneck module, while the other segment goes through feature extraction via multiple ParNet bottleneck modules. The resulting feature maps are then merged, and the final results are produced by the second convolutional layer.



**FIGURE 4** Improved network module for the neck.

In the `forward_split()` function, the `split()` function is utilized instead of the `chunk()` function to streamline the feature splitting process. This results in C2f-ParNet reducing computational complexity and improving feature extraction efficiency. Moreover, it helps in addressing false alarms and missed detections that are often caused by target occlusion.

### 2.3.2 | Improved CARAFE module

CARAFE [27] is a lightweight, general-purpose sampling operator in deep learning. Its primary function is to apply adaptive convolution to the features extracted by the backbone network, effectively narrowing the region of interest. This capability allows CARAFE to improve the fusion of target features, thereby enhancing the model's overall performance. By incorporating the CARAFE operator, the model can flexibly adjust convolution operations, selectively concentrating on areas of interest, which consequently enhances sensitivity to targets. This lightweight sampling operator is critical for improving model performance, particularly in the domains of target perception and feature fusion.

At each position, CARAFE can leverage low-level contextual information to predict a reassembly kernel and reassemble features within a predefined neighbourhood. Its operation primarily involves two steps: the first is to predict a reassembly kernel based on the content at each target's position, and the second step is to reassemble the features using the predicted kernel. The CARAFE structure is illustrated in Figure 5. Given a feature map  $x$  with dimensions  $C \times H \times W$  and an upsampling ratio, CARAFE generates a new feature map  $x_0$ . The dimensions of the new feature map  $x_0$  are  $C \times \delta H \times \delta W$ , and the formulas for calculating the position-wise kernel and reassembly kernel are as follows:

$$W_{l'} = \psi(N(x_l, k_{\text{encode}})) \quad (1)$$

$$x'_{l'} = \varphi(N, (x_l, k_{\text{up}}), W_{l'}) \quad (2)$$

where  $\psi$  represents the kernel prediction module,  $l'$  denotes the corresponding kernel positions, and  $\varphi$  signifies the reassembly of  $x_l$  with the kernel  $W_{l'}$  to obtain  $x'_{l'}$ .

In Figure 5, to minimize channel parameters and computational costs of the feature map, while enhancing the efficiency



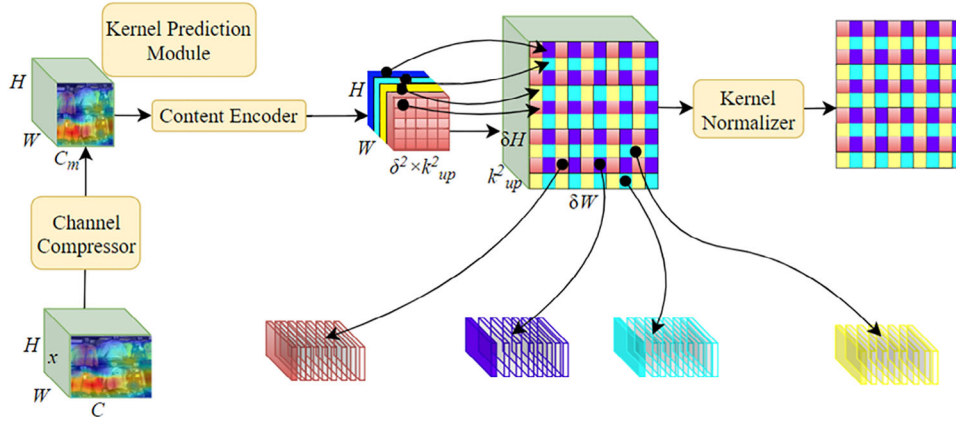


FIGURE 5 The overall framework of CARAFE.

of feature fusion within the neck network, the new feature map  $x_0$  undergoes channel compression to generate the kernel prediction module  $C_m \times H \times W$ , where  $C_m$  represents the value compressed using a  $1 \times 1$  convolutional layer. The kernel prediction module is primarily responsible for generating reassembly kernels in a content-aware manner. The source positions on feature map  $x$  correspond to the two target positions on the up sampled feature map  $x_0$  with a factor of  $\delta^2$ , and each target position requires a reassembly kernel  $k_{up}^2$ , the size of the reassembly kernel is denoted by  $k_{up}$ . Therefore, the size of the new feature map is  $(\delta^2 k_{up}^2) \times H \times W$ , under the same budget.

CARAFE initially utilizes the input feature map to predict a position-specific upsampling kernel, normalizing this kernel accordingly. Subsequently, it reorganizes features based on the predicted upsampling kernel, which is then mapped back to the input feature map for each position in the output feature map. Notably, different channels at the same position share the same upsampling kernel.

### 2.3.3 | Improvement of the IoU

In supermarket object detection, the YOLOv8n model utilizes the complete intersection over union (CIoU) function as the regression loss function for bounding box prediction, thereby estimating the model's detection performance. Here,  $w$  and  $b$  represent the width and height of the predicted bounding box, while  $w^r$  and  $b^r$  represent the width and height of the true bounding box. The coordinates of the bounding box centres for the predicted and true boxes are denoted by  $s$  and  $s^r$  respectively, and  $d$  represents the Euclidean distance between  $s$  and  $s^r$ . Additionally,  $w^m$  and  $b^m$  represent the width and height of the minimum enclosing rectangle for the predicted and true boxes. IoU represents the Intersection over Union. The CIoU loss formula is expressed as follows:

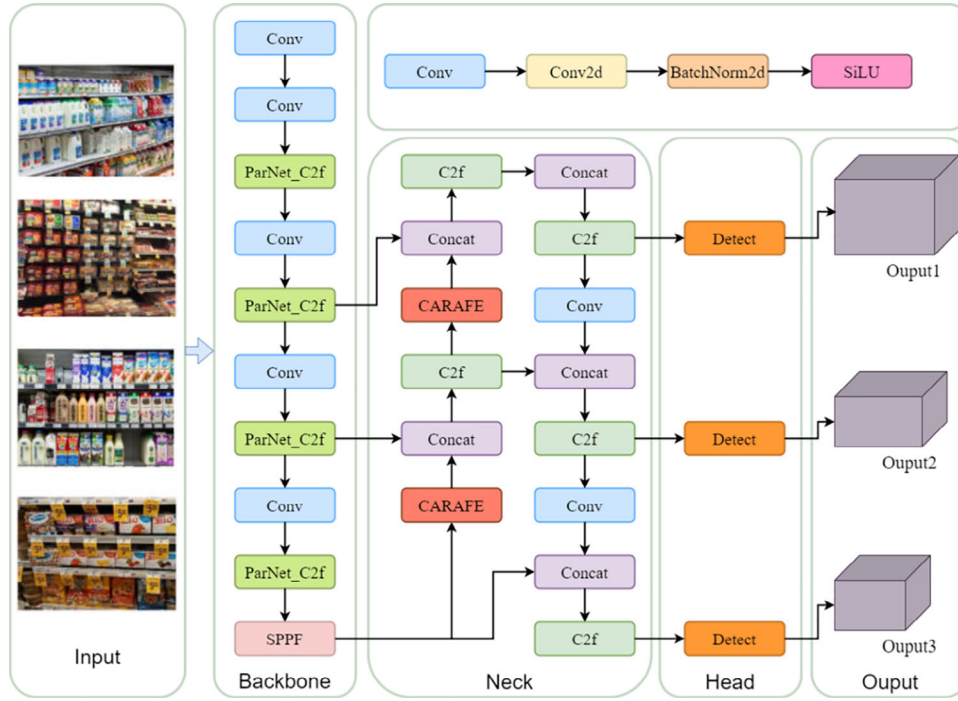
$$L_{CIoU} = 1 - IOU + \frac{d^2(s, s^r)}{(w^m)^2 + (b^m)^2} + uv \quad (3)$$

$$\text{Where: } u = \frac{v}{(1 - IOU) + v}; v = \frac{4}{\pi^2} \left( \arctan \frac{w^r}{b^r} - \arctan \frac{w}{b} \right)^2 \quad (4)$$

The complete IoU (CIoU) metric introduces several enhancements over traditional IoU, incorporating boundary regression distance, centre point offset, overlap area, and the comparison of predicted versus actual width-height ratios. Unlike DIOU, which cannot distinguish cases where bounding box centre points are coincident, CIoU improves boundary convergence rates by addressing this limitation. Nevertheless, CIoU does not fully rectify the alignment discrepancies between predicted and actual bounding boxes. As delineated in ref. [30], the parameter ' $v$ ' within Equation (4) quantifies the disparity in aspect ratios rather than representing the precise width and height dimensions of the bounding boxes. Furthermore, the presence of low-quality data samples significantly exacerbates regression loss issues, presenting substantial challenges in optimizing the data, as discussed in ref. [28].

To mitigate the imbalance between high and low-quality sample data in bounding box regression (BBR), Zhang et al. [28] introduces the focal extended intersection over union version 1 (Focal EIoUv1). This method attempts to rectify the disparity in sample quality; however, it employs a static focusing mechanism (FM) that does not fully harness its monotonic potential. In contrast, the weighted intersection over union (WIoU) approach addresses this limitation by implementing a dynamic, non-monotonic focusing mechanism, significantly augmenting the algorithm's performance as reported in studies [29–30]. WIoU constructs a dual-layer attention mechanism to enhance the model's generalization ability. WIoU can be divided into WIoUv1, WIoUv2, and WIoUv3 [31], and comprehensive later-stage experiments prove that WIoUv3 exhibits strong performance. Therefore, the system model chooses WIoUv3 as the bounding box regression loss function, while the formula for WIoUv1 is as follows:

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \quad (5)$$



**FIGURE 6** The improved YOLOv8n network structure.

The calculation for variables  $R_{IoU}$  and  $L_{IoU}$  in the equation is done as per Equations (6) and (7):

$$R_{WIoU} = \exp\left(\frac{(x - x')^2 + (y - y')^2}{((W^M)^2 + (H^M)^2)}\right) \quad (6)$$

$$L_{IoU} = 1 - IoU \quad (7)$$

where  $x$ ,  $y$ ,  $x'$ , and  $y'$  respectively represent the centre coordinates of the predicted box and the true box;  $W^M$  and  $H^M$  represent the width and height of the true box.

In order to avoid the problem of low sample quality caused by the blurred deformation of images due to the movement of shopping robots,  $WIoUv3$  is constructed using  $\beta$  and  $WIoUv1$ . The formula is as follows:

$$L_{WIoUv3} = \frac{\beta}{\lambda \alpha^{\beta-\lambda}} L_{WIoUv1} \quad (8)$$

$$\beta = L_{IoU}^* / L_{IoU} \in [0, +\infty); \quad (9)$$

In summary, the improved YOLOv8n architecture aims to mitigate issues of false positives and false negatives in target detection by shopping robots, influenced by diverse factors. This is achieved by embedding the ParNetAttention within the C2f module of the backbone network and by integrating the CARAFE module into the neck network. The loss function employed is  $WIoUv3$ . The refined architectural framework of YOLOv8n is shown in Figure 6.

**TABLE 2** Hardware and environment parameters.

Operating systems	Ubuntu 18.04.5 LTS
Epochs	150
CPU	Intel(R) Xeon(R) Platinum 8350C
GPU	NVIDIA GeForce RTX 3090
$\lambda$ (WIoUv3)	1.9
$\alpha$ (WIoUv3)	3
GPU memory size	24 GB
Deep learning architecture	Python-3.8.17-torch-2.0.0
Batch size of initiation	32
Initial learning rate	0.01

## 2.4 | Training and evaluation

To fully validate the effectiveness of the algorithm in later experiments, the hardware platform and environmental parameters utilized during the experimental training, testing, and validation stages are presented in Table 2.

During the experiment, performance metrics such as precision and recall can be employed to assess the accuracy of the model's detection. However, excessively high precision and recall may result in issues of model omission and false alarms, making them less suitable as crucial model evaluation metrics. Therefore, the system prioritizes mean average precision (mAP) and confidence score F1 as essential performance metrics for the algorithm.

Precision is quantified as the ratio of correctly identified positives to the total predicted positives, as formulated in



Equation (10). Recall measures the ratio of correctly predicted positives to all actual positives, assessing the model's detection coverage, delineated in Equation (11). Moreover, the mean average precision (mAP) serves as an essential metric for evaluating model performance in object detection, capturing both accuracy and comprehensiveness across various target categories, as specified in Equation (12). The F1 score, integrating precision and recall, offers a comprehensive evaluation of the network's overall efficacy, as explicated in Equation (13).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (10)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (11)$$

$$\text{mAP} = \frac{\sum_{i=1}^N \int_0^1 \text{Precision}(\text{Recall}) d(\text{Recall})}{N} \quad (12)$$

$$F_1 = 2 \times \text{Precision} \times \frac{\text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

In these equations, TP represents true positives, FP represents false positives, and FN represents false negatives. Specifically for object detection algorithms, the average precision within the IoU range of 0.5 is denoted as mAP@50, while the average precision within the broader IoU range from 0.5 to 0.95 is denoted as mAP@50-95.

### 3 | EXPERIMENTAL RESULTS

#### 3.1 | Comparative experiment for loss function

To ascertain the efficacy of the novel loss function WIoUv3, a comparative analysis was undertaken against existing metrics, including CIoU, DIoU, GIoU [32], SIOU [33], WIoUv1, and WIoUv2. The experimental framework was standardized using the YOLOv8n loss function, and the evaluation was conducted over 150 training epochs. The resultant training loss trajectory, illustrated in Figure 7a, demonstrates a marked initial decline in the WIoUv3 loss curve, followed by a plateau, indicative of enhanced optimization efficiency. Notably, WIoUv3 consistently maintains lower loss values relative to the comparative loss functions, signifying improved predictive accuracy within the detection model. Figure 7b delineates the mean average precision at 50% Intersection over Union (mAP50) training curve for WIoUv3, which exhibits a swift initial ascent before reaching stability, outperforming alternative functions in terms of detection capabilities.

And Table 3 further elucidates these findings, presenting mAP@50 values where GIoU records the lowest at 58.0% and 37.9%, whereas the baseline model utilizing CIoU reflects slightly higher values at 60.0% and 39.5%. In contrast, WIoUv3 achieves mAP50 values of 61.2% and 40.1%, marking a 1.2% enhancement over the baseline. Moreover, when juxtaposed with other evaluated functions (DIoU, GIoU, SIOU,

WIoUv1, and WIoUv2), the integration of YOLOv8n with WIoUv3 attains the highest F1 score in Table 3, representing a 1.9% increment over the baseline configured with CIoU. This substantiates the premise that WIoUv3 confers a significant improvement in the model's detection performance.

#### 3.2 | Ablation experiments

In the later stages of the experiment, to assess the impact of different enhancement strategies on the original network, seven sets of ablation experiments are conducted independently for the improved modules. The experimental results are presented in Table 4, where "CPN" denotes the C2f-ParNet module, "C" represents the CARAFE module, and "W" represents the WIoUv3 module.

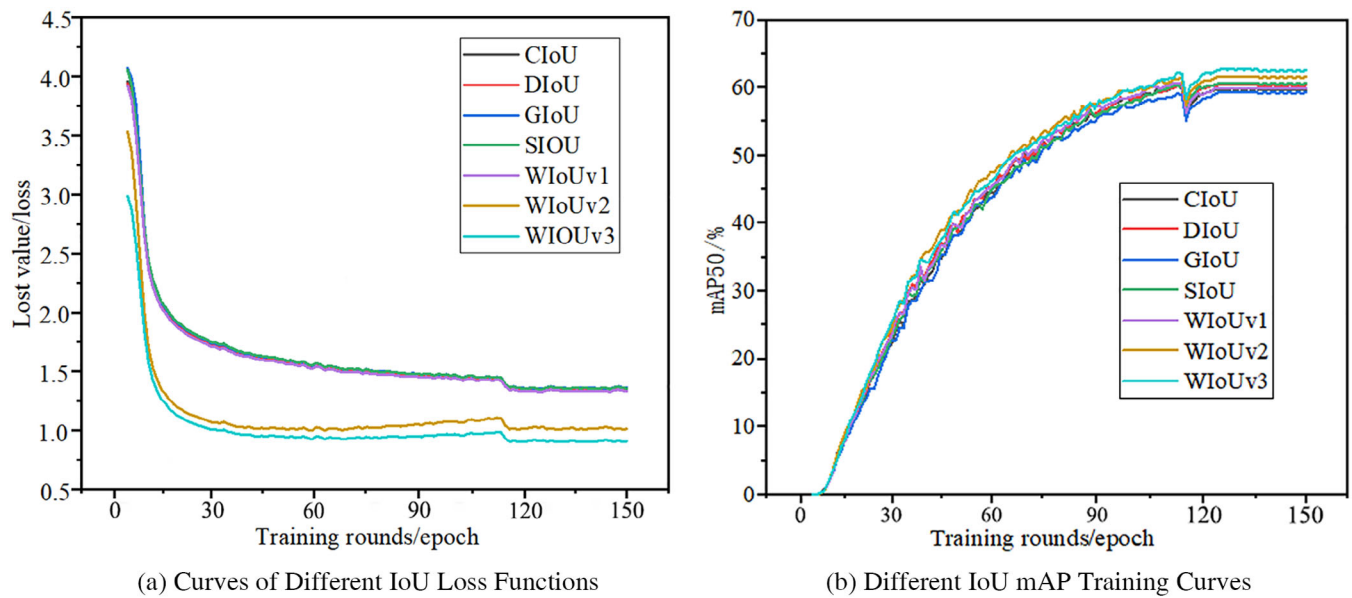
Table 4 illustrates that the mAP@50 of the upgraded YOLOv8n+CPN+W+C algorithm achieved 65.9%, representing a 5.9% improvement over the baseline model. Moreover, the mAP@50-95 exhibited a gain of 3.7 percentage points compared to the baseline. In comparison to other models, the enhanced algorithm displayed increases in mAP@50 of 3.5%, 4.7%, 3.7%, 1.4%, 1.1%, and 2.2%, respectively. The F1 score peaked at 63.5%, signifying a notable enhancement in network model stability. The frames per second (FPS) reached its peak performance at 106.5 FPS. Despite a slight decrease in Precision due to the increased network layers post-enhancements, it still outperformed the majority of shopping robot object detection models.

The Locount dataset's extensive volume predisposes the algorithmic model to potential overfitting due to the generation of redundant information. To mitigate this, the model was evaluated using a preprocessed version of the dataset. The training and validation trajectories are depicted in Figure 8. For the initial 100 epochs, the loss curves in Figure 8a–h exhibits a rapid decrease followed by a more gradual reduction, stabilizing between epochs 100 and 150. Conversely, in Figure 8d,e,i,j, the training and validation curves initially rise sharply in the first 100 epochs, followed by a slower increase, with the trajectory in Figure 8d achieving stabilization post-epoch 100.

Considering the prevalence of small and occluded objects in the dataset, along with its substantial size, Figure 8e,i,j exhibits a slight decrease followed by a stable state in the testing and validation curves. Consequently, the experimental results indicate that the enhanced algorithm model does not encounter overfitting issues during the training and validation processes, demonstrating favourable convergence properties.

#### 3.3 | Comparative experiment of models

To underscore the marked superiority and efficacy of the augmented algorithm, a series of comparative experiments were systematically orchestrated. As detailed in Table 5, established algorithms such as SSD, RetinaNet [34], Cascade R-CNN [35], and Faster R-CNN are characterized by significant computational overheads, thereby limiting their utility in real-time



**FIGURE 7** Curves of different IoU training results.

**TABLE 3** Performance comparison of the different loss functions.

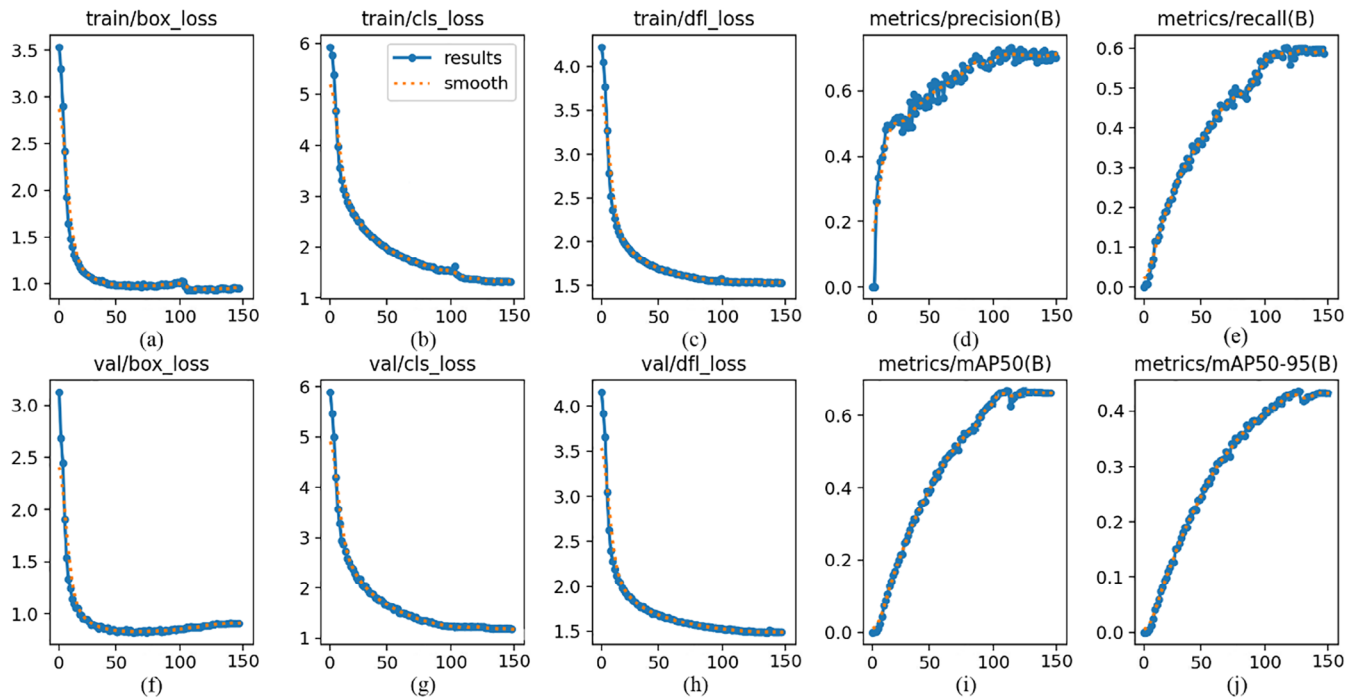
Loss function	Precision/%	Recall/%	F1/%	mAP50/%	mAP50-95/%
CIOU (baseline)	70.8	50.2	58.7	60.0	39.5
DIOU	67.6	52.9	59.4	58.7	39.0
GIOU	67.0	52.6	58.9	58.0	37.9
SIOU	70.3	50.8	59.0	58.1	38.2
WIOUv1	70.3	52.0	59.8	59.3	39.3
WIOUv2	71.3	52.5	60.5	61.0	40.0
WIOUv3	<b>71.7</b>	52.5	<b>60.6</b>	<b>61.2</b>	<b>40.1</b>

**TABLE 4** Results of ablation experiments.

Models	Precision/%	Recall/%	F1/%	mAP50/%	mAP50-95/%	Frame rate/FPS
Yolov8n(baseline)	70.8	50.2	58.7	60.0	39.5	78.7
YOLOv8n+CPN	71.9	54.4	61.9	62.4	41.2	101.1
YOLOv8n+W	71.7	52.5	60.6	61.2	40.1	76.3
YOLOv8n+C	71.6	55.4	62.5	62.2	41.0	79.4
YOLOv8n+CPN+W	74.6	54.2	64.3	64.5	41.8	102.5
YOLOv8n+CPN+C	71.6	56.0	62.8	64.8	42.8	105.7
YOLOv8n+W+C	70.7	56.0	62.5	63.7	41.7	101.6
YOLOv8n+CPN+W+C	70.2	<b>57.9</b>	<b>63.5</b>	<b>65.9</b>	<b>43.2</b>	106.5

object detection contexts. Conversely, models like YOLOv5, YOLOv7, and YOLOv8n, although less computationally intensive, suffer from reduced detection accuracy. In this study, mAP is used to evaluate model performance. This is due to the nature of object detection tasks, which typically rely on the predictions obtained after non-maximum suppression (NMS). A high recall

may lead to a significant number of false positives, resulting in many predicted bounding boxes with low precision remaining after NMS. Conversely, a high precision might indicate a substantial number of missed detections, leading to too few predicted boxes, although the accuracy within these limited boxes may be relatively high.



**FIGURE 8** Training and validation curves of the YOLOv8-CPN-CW model.

**TABLE 5** Comparison of the experimental results with various algorithms.

Models	Precision/%	Recall/%	F1/%	mAP50/%	Computational volume/GFLOPs
SSD	62.3	43.5	51.2	54.8	35.8
RetinaNet	60.2	40.1	48.1	51.7	254
Cascade R-CNN	65.6	44.7	53.2	58.6	238.7
Faster R-CNN	61.3	42.2	50.0	53.8	267.3
Yolov5s	63.8	46.6	53.9	56.4	18.3
Yolov7-tiny	69.7	48.3	57.1	57.2	14.8
Yolov8n	70.8	50.2	58.7	60.0	13.1
Ours	70.2	57.9	63.5	65.9	12.9

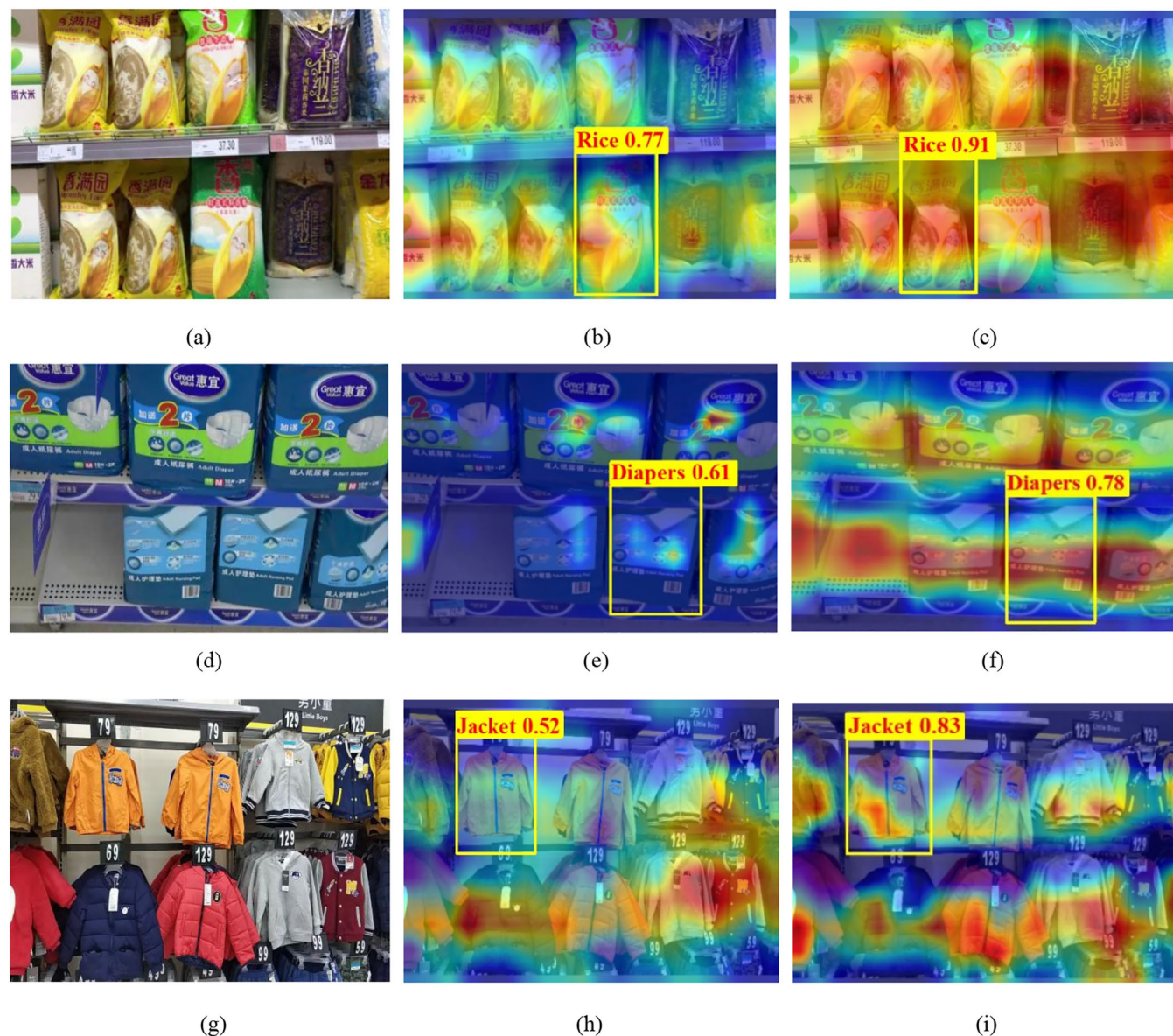
The novel YOLOv8-CPN-CW algorithm achieves a computational burden of merely 12.9 GFLOPs, thereby facilitating lightweight computation while simultaneously preserving robust performance metrics, as evidenced by an mAP@50 value. This algorithm outperforms its counterparts by improving performance by 14.2% over the lowest benchmarked value and by 5.9% over the previously highest value observed. Additionally, the  $F1$  score of the enhanced model peaks at 63.5%, representing a 4.7% improvement over the baseline model. These experimental results emphatically demonstrate that the refined YOLOv8-CPN-CW algorithm not only achieves the highest detection efficiency among the tested popular object detection algorithms but also demands lower computational resources. This translates to diminished operational costs, streamlined deployment, and overall enhanced efficiency, making it particularly advantageous for real-time applications.

### 3.4 | Visualization analysis

Recognizing the inherent challenges in interpreting deep learning models, this paper undertakes a comprehensive analysis of our model's visual detection capabilities, employing both heatmaps and confusion matrices. To evaluate the robustness and generalization of our approach, real-world validations on the preprocessed Locount dataset, both before and after implementing enhancements are conducted. The focus was particularly directed towards challenging scenarios, including the detection of multiple overlapping objects, objects with similar colour profiles, and images affected by motion blur.

This study utilizes gradient-weighted class activation mapping (Grad CAM) [36] to generate insightful visual heatmaps for both the baseline YOLOv8n model and its advanced iteration, YOLOv8-CPN-CW. Grad CAM facilitates an intuitive





**FIGURE 9** Original image and visualized heatmap. (a,d,g) Original image. (b,e,h) Heatmap before improvement. (c,f,i) Heatmap after improvement.

visualization of feature extraction by computing gradient values through the backpropagation of confidence scores from the model's output classes, as illustrated in Figure 9. In these visualizations, areas of higher model focus are highlighted in deep red and yellow, indicating regions with higher gradient values, whereas areas of lesser focus are shown in blue and cyan.

Figure 9 presents a series of comparative heatmaps: Figure 9a,d,g displays the original scenes; Figure 9b,e,h shows the heatmaps generated by the standard YOLOv8n model; and Figure 9c,f,i depicts the heatmaps from the enhanced YOLOv8-CPN-CW model. A direct comparison between Figure 9b,c showcases a discernible intensification of deep red and yellow regions in Figure 9c, signalling a boost in the confidence levels detected by the model. This enhancement in heatmap

intensity not only demonstrates the improved model's effectiveness but also substantiates that the YOLOv8-CPN-CW model's attention mechanisms are more accurately focused on the critical centres of target objects, thereby yielding more precise bounding box predictions.

Following a series of experiments encompassing loss function comparisons, ablation studies on algorithm enhancements, comparative assessments with other algorithms, and visual analyses, a pivotal phase in validating the practical efficacy of the enhanced YOLOv8-CPN-CW algorithm involves random validation within real-world scene detection results using the Locount dataset. This dataset contains images featuring objects of similar colours, instances of multiple object occlusions, and motion blur (Figure 10).





**FIGURE 10** Comparison of real-world scenes between YOLOv8n and the improved algorithm. (a,b) For the same colour target, the actual detection results of YOLOv8n and the improved algorithm. (c,d) For multiple target occlusion, the actual detection results of YOLOv8n and the improved algorithm. (e,f) For motion-blurred images, the actual detection results of YOLOv8n and the improved algorithm.

## 4 | CONCLUSIONS AND DISCUSSIONS

Due to challenges such as occlusions of multiple targets, similar colours, and motion blur, the shopping robot faces significant

issues with false positives and false negatives. To address these challenges, we propose a modified YOLOv8-CPN-CW algorithm. This algorithm enhances the lightweight YOLOv8n network model by incorporating the C2f-ParNet module into

the backbone network and embedding the CARAFE module into the neck network. The loss function utilizes WIoUv3 to evaluate target category correctness and the presence of predicted targets. Comparative experiments, including comparisons of loss functions, ablation studies, and comparisons with other algorithms, are conducted on the improved model. The experimental results confirm the effectiveness of the proposed algorithm. In conclusion, the study demonstrates the efficacy of the modified algorithm.

- To confirm the superiority of the WIoUv3 loss function, experiments are conducted to compare the performance of CIoU, DIoU, GIoU, SIOU, WIoUv1, WIoUv2, and WIoUv3. The experimental results reveal that the loss curve of WIoUv3 follows the optimal trend, with mAP50 and F1 values of 0.612 and 0.606, respectively, both reaching their peak.
- The network model, utilizing the WIoUv3 loss function, underwent ablation experiments involving the integration of the C2f-ParNet module into the backbone network of the YOLOv8n model, and the embedding of the CARAFE module into the neck. In comparison to the original YOLOv8n algorithm, the improved YOLOv8-CPN-CW algorithm demonstrated a 4.7% increase in mAP@50 and a 2.9% increase in the F1 score. Furthermore, when the C2f-ParNet module and CARAFE module are individually added, the mAP@50 improved by 3.3% and 2.5%, respectively. The improved algorithm not only boosts detection performance but also achieves lightweight computation, thereby enhancing overall efficiency.
- To assess the practical effectiveness of the improved algorithm in real-world scenarios, real-world testing experiments were conducted on the YOLOv8-CPN-CW algorithm, taking into account factors such as target occlusion, similar colours, and motion blur. The experimental results exhibited substantial improvements over the baseline YOLOv8n algorithm. The confidence values for the selected bounding boxes were notably enhanced, and both false positive and false negative rates were reduced.

This study introduces several strategic and creative improvements to the YOLOv8n algorithm aimed at enhancing its application in robotic vision systems, specifically for shopping robots. First, this study integrates the C2f module into the ParNetAttention within the backbone network. This addition is designed to augment the algorithm's capability to discriminate finer details in complex visual scenes, thereby enhancing overall detection accuracy. Secondly, the replacement of the conventional Upsample module with the advanced CARAFE module in the neck network significantly improves the model's perceptual ability. This change allows for a more refined reconstruction of image details, facilitating better accuracy in object detection tasks. Furthermore, it adopts the WIoUv3 loss function to refine the precision of bounding box estimations, directly contributing to improvements in the average accuracy of object detection. This alteration is particularly beneficial for

the nuanced requirements of shopping environments where precise object recognition is crucial.

The suggested further improvements of the proposed model are as follows: (1) Due to the addition of some modules, the network model becomes more complex as a consequence of the increase in network size. The algorithm can be effectively pruned at a later stage to reduce the size of the network storage and deploy the algorithm efficiently while ensuring that the accuracy of the object detection is slightly changed; (2) Further applications with different datasets on other fields can be executed as well [37–39]. Fine-tuning the model on a domain-specific dataset could enhance its ability to detect relevant features and improve overall performance; (3) For the problem of insufficient precision and recall, there are some possible strategies such as expanding the training dataset through various augmentation techniques, a thorough investigation of hyperparameter tuning and improving the NMS via incorporating additional post-processing steps that can reduce false positives while maintaining a higher number of true positives. (4) We plan to incorporate model detection testing on the robot hardware to assess how the algorithm performs in dynamic environments and under practical constraints. We will also incorporate the need for model pruning and quantization to achieve model lightweight that can significantly accelerate inference speed.

## AUTHOR CONTRIBUTIONS

**Yawen Zhao:** Conceptualization; formal analysis; investigation; methodology; resources; software; visualization; writing—original draft. **Defu Yang:** Conceptualization; funding acquisition; methodology; resources. **Sheng Cao:** Funding acquisition; resources; writing—review and editing. **Bingyu Chai:** Funding acquisition; methodology; resources. **Maryamah:** Supervision; visualization. **Mahmud Iwan Solihin:** Conceptualization; formal analysis; investigation; methodology; supervision; writing—review and editing.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The dataset is available upon request.

## ORCID

Mahmud Iwan Solihin  <https://orcid.org/0000-0002-5293-7466>

## REFERENCES

1. Qian, W., Hu, C., Wang, H., Lu, L., Shi, Z.: A novel target detection and localization method in indoor environment for mobile robot based on improved YOLOv5. *Multimedia Tools Appl.* 82(18), 28643–28668 (2023)
2. Ang, K., Ang, K.M., Juhari, M.R.B.M., Wong, C.H., Sharma, A., Ang, C.K., Lim, W.H.: Classification of wafer defects with optimized deep learning model. In: *Proceeding of 28<sup>th</sup> International Conference on Artificial Life and Robotics*, pp. 609–614. IEEE, Piscataway, NJ (2023)
3. Cheng, W.L., Pan, L., Juhari, M.R.B.M., Sharma, A., Rahman, H., Ang, C.K., Tiang, S.S., Lim, W.H.: Multi chaotic flow direction algorithm for feature selection. In: *Proceedings of 28<sup>th</sup> International Conference on Artificial Life and Robotics*, pp. 599–604. IEEE, Piscataway, NJ (2023)



4. Cheng, W.L., Pan, L., Juhari, M.R.B.M., Wong, C.H., Sharma, A., Lim, T.H., Tiang, S.S., Lim, W.H.: Chaotic african vultures optimization algorithm for feature selection. In: Proceedings of 28<sup>th</sup> International Conference on Artificial Life and Robotics, pp. 593–598. IEEE, Piscataway, NJ (2023)
5. Liu, Z., Wang, J., Li, J., Liu, P., Ren, K.: A novel multiple targets detection method for service robots in the indoor complex scenes. *Intell. Serv. Rob.* 16(4), 453–469 (2023)
6. Meng, Q., Yang, J., Zhang, Y., Yang, Y., Song, J., Wang, J.: A robot system for rapid and intelligent bridge damage inspection based on deep-learning algorithms. *J. Perform. Constr. Facil.* 37(6), 04023052 (2023)
7. Jiang, J., Ying, S., Fu, W., Jiang, X.: Structure design and system implementation of a supermarket shopping robot based on deep learning. *Int. J. Data Sci.* 8(1), 1–15 (2023)
8. Zhang, X., Lu, H., Xu, Q., Peng, X., Li, Y., Liu, L., Zhang, W.: Image recognition of supermarket shopping robot based on CNN. In: Proceeding of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 1363–1368. IEEE, Piscataway, NJ (2020)
9. Aydın, M., Erdemir, G.: An object detection and identification system for a mobile robot control. *Balkan J. Electr. Comput. Eng.* 5(2), 73–76 (2017)
10. Jiang, L., Nie, W., Zhu, J., Gao, X., Lei, B.: Lightweight object detection network model suitable for indoor mobile robots. *J. Mech. Sci. Technol.* 36(2), 907–920 (2020)
11. Zhang, L.J., Fang, J.J., Liu, Y.X., Le, H.F., Rao, Z.Q., Zhao, J.X.: CR-YOLOv8: Multiscale object detection in traffic sign images. *IEEE Access* 12, 219–228 (2023)
12. Barba-Guaman, L., Naranjo, J.E., Ortiz, A., Gonzalez, J.G.P.: Object detection in rural roads through SSD and YOLO framework. In: Proceedings of the Trends and Applications in Information Systems and Technologies, pp. 176–185. Springer, Cham (2021)
13. Kim, J.A., Sung, J.Y., Park, S.H.: Comparison of Faster-RCNN, YOLO, and SSD for real-time vehicle type recognition. In: Proceeding of IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), pp. 1–4. IEEE, Piscataway, NJ (2020)
14. Gong, C., Li, A., Song, Y., Xu, N., He, W.: Traffic sign recognition based on the YOLOv3 algorithm. *Sensors* 22(23), 9345 (2022)
15. Liu, J., Wang, X., Miao, W., Liu, G.: Tomato pest recognition algorithm based on improved YOLOv4. *Front. Plant Sci.* 13, 814681 (2022)
16. Liu, H., Duan, X., Lou, H., Gu, J., Chen, H., Bi, L.: Improved GBS-YOLOv5 algorithm based on YOLOv5 applied to UAV intelligent traffic. *Sci. Rep.* 13(1), 9577 (2023)
17. Zhu, Q., Ma, K., Wang, Z., Shi, P.: YOLOv7-CSAW for maritime target detection. *Front. Neurorob.* 17, 1210470 (2023)
18. Talaat, F.M., ZainEldin, H.: An improved fire detection approach based on YOLO-v8 for smart cities. *Neural Comput. Appl.* 35(28), 20939–20954 (2023)
19. Follmann, P., Böttger, T., Härtinger, P., König, R., Ulrich, U.: MVTec D2S: Densely segmented supermarket dataset. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 569–585. Springer-Verlag, Berlin, Heidelberg (2018). <https://doi.org/10.48550/arXiv.1804.08292>
20. Wei, X.S., Cui, Q., Yang, L., Wang, P., Liu, L.Q.: A large-scale retail product checkout dataset. In: Proceedings of the Computer Vision and Pattern Recognition, pp. 1–24. IEEE, Piscataway, NJ (2019). <https://doi.org/10.48550/arXiv.1901.07249>
21. Hao, Y., Fu, Y.W., Jiang, Y.G.: Take goods from shelves: A dataset for class-incremental object detection. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval, pp. 271–278. ACM, New York, NY (2019)
22. Goldman, E., Herzig, R., Eisenschtat, A., Ratzon, O., Levi, I., Goldberger, J., Hassner, T.: Precise detection in densely packed scenes. In: Proceeding of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5227–5236. IEEE, Piscataway, NJ (2019)
23. Cai, Y.Q., Wen, L.Y., Zhang, L., Du, D.W., Wang, W.Q.: Rethinking object detection in retail stores. *Proc. AAAI Conf. Artif. Intell.* 35(2), 947–954 (2021)
24. Lin, T.Y., Dollár, P., Girshick, R., He, K.M., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the Computer Vision and Pattern Recognition, pp. 2117–2125. IEEE, Piscataway, NJ (2017)
25. Liu, S., Qi, L., Qin, H.F., Shi, J.P., Jia, J.Y.: Path aggregation network for instance segmentation. In: Proceedings of the Computer Vision and Pattern Recognition, pp. 8759–8768. IEEE, Piscataway, NJ (2018)
26. Goyal, A., Bochkovskiy, A., Deng, J., Koltun, V.: Non-deep networks. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 6789–680. Curran Associates Inc., Red Hook, NY (2022)
27. Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D.: Carafe++: Unified content-aware reassembly of features. *IEEE Trans. Pattern Anal. Mach. Intell.* 44(9), 4674–4687 (2021)
28. Zhang, Y.F., Ren, W., Zhang, Z., Jia, Z., Wang, L., Tan, T.: Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* 506, 146–157 (2022)
29. Tong, Z., Chen, Y., Xu, Z., Yu, R.: Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv:2301.10051* (2023). <https://arxiv.org/abs/2301.10051>
30. Yuan, H.C., Tao, L.: Detection and identification of fish in electronic monitoring data of commercial fishing vessels based on improved Yolov8. *J. Dalian Ocean Univ.* 38(3), 533–542 (2023)
31. Cho, Y.J.: Weighted intersection over union (wIoU): A new evaluation metric for image segmentation. *arXiv:2107.09858* (2021). <https://arxiv.org/abs/2107.09858>
32. Hou, Z., Liu, X., Chen, L.: Object detection algorithm for improving non-maximum suppression using GIoU. *IOP Conf. Ser.: Mater. Sci. Eng.* 790, 012062 (2020)
33. Shen, S., Zhang, X., Yan, W., Xie, S., Yu, B., Wang, S.: An improved UAV target detection algorithm based on ASFF-YOLOv5s. *Math. Biosci. Eng.* 20(6), 10773–10789 (2023)
34. Ale, L., Zhang, N., Li, L.: Road damage detection using RetinaNet. In: Proceeding of the 2018 IEEE International Conference on Big Data (Big Data), pp. 5197–5200. IEEE, Piscataway, NJ (2018)
35. Qi, L., Li, B., Chen, L., Wang, W., Dong, L., Jia, X., Wang, D.: Ship target detection algorithm based on improved faster R-CNN. *Electronics* 8(9), 959 (2019)
36. Long, Y., Yang, Z.Y., He, M.F.: Fruit-thinning phase apple target detection method based on improved YOLOv7. *Trans. Chin. Soc. Agric. Eng.* 39(14), 191–199 (2023)
37. Li, W., Solihin, M.I., Nugroho, H.A.: RCA: YOLOv8-based surface defects detection on the inner wall of cylindrical high-precision parts. *Arabian J. Sci. Eng.* 49, 12771–12789 (2024)
38. Yang, D., Solihin, M.I., Zhao, Y., Yao, B., Chen, C., Cai, B., Machmudah, A.: A review of intelligent ship marine object detection based on RGB camera. *IET Image Proc.* 18(2), 281–297 (2024)
39. Li, G., Xie, Y., Lu, Y., et al.: Enhancing precision object detection and identification for autonomous vehicles through YOLOv5 refinement with YOLO-ALPHA. *Proc. Int. Conf. Artif. Life Rob.* 2, 889–894 (2024)

**How to cite this article:** Zhao, Y., Yang, D., Cao, S., Cai, B., Maryamah, M., Solihin, M.I.: Object detection in smart indoor shopping using an enhanced YOLOv8n algorithm. *IET Image Process.* 18, 4745–4759 (2024). <https://doi.org/10.1049/ipr2.13284>