scientific reports

Check for updates

OPEN

Lightweight object detection model for food freezer warehouses

Jiayu Yang, Zhihong Liang✉, Mingming Qin, Xingyu Tong, Fei Xiong & Hao An

Warehouses are critical logistics nodes, with food freezer warehouses playing a key role in ensuring food quality while facing challenges such as high-density item distribution and extremely low temperatures required for occupational safety. Traditional management methods struggle to meet these demands, underscoring the need for intelligent and digital solutions to improve efficiency and mitigate safety risks. This study proposes the YOLOv8-RSS model, a lightweight and high-precision approach tailored for food freezer warehouse scenarios. The model incorporates the novel C2f_RDB module, which enhances detection accuracy while reducing parameter count and computational load. Additionally, the SimAM attention mechanism is applied to the Backbone's final layer, enabling focus on critical image information without increasing parameters. Soft-NMS replaces the traditional NMS method, further improving detection accuracy. Experiments conducted on the food freezer warehouse dataset demonstrate that the YOLOv8-RSS model reduced the parameter count by 0.05 M, decreased FLOPs by 0.8G, increased mAP@0.5 by 1.4%, and improved mAP@0.5:0.95 by 3.9%. The YOLOv8-RSS is designed to meet the complex detection demands in food freezer warehouses, enabling precise and rapid detection of personnel and forklifts. It provides strong technical support for addressing various challenges in these environments and holds significant application value.

Keywords Deep learning, Object detection, Warehouse Management, You-only-look-once (YOLO)

Warehouses are the core nodes of logistics systems, playing a pivotal role in critical functions such as storage, sorting, and dispatching, which significantly impact logistics operations' efficiency and quality. Traditional warehouse management methods, often reliant on manual inventory checks and scheduling, work effectively in low-complexity scenarios but reveal limitations in addressing the growing complexities of modern logistics. On the one hand, the expansion of warehouse-scale and the diversification of stored goods have significantly increased management complexity, making it challenging for manual operations to provide real-time, accurate inventory tracking and dynamic task monitoring. This often results in inefficiencies and resource waste[1,2]. On the other hand, the rising number of forklifts and personnel and their unpredictable activities further escalate management challenges and introduce safety risks. These issues are particularly pronounced in specialized environments such as freezer warehouses, where extremely low temperatures pose significant safety concerns for workers, exacerbating warehouse management challenges.

In this context, enhancing the intelligence and informatization of warehouse management has become a pressing need. Computer vision technologies offer a viable solution for this transformation. By leveraging object detection, critical information can be swiftly extracted from complex visual data, enabling precise detection of personnel and forklifts, thereby improving warehouse monitoring efficiency and ensuring operational safety[3].

One of the main problems in computer vision is object detection[4], which has several applications in pedestrian detection[5], face detection[6], and autonomous driving[7]. Object detection algorithms have achieved significant breakthroughs recently, motivated by deep learning technologies' broad adoption and quick development. The continuous improvement in their accuracy and efficiency is powerfully driving technological innovation and expanding the scope of applications across various industries.

There are two types of deep learning-based object detection algorithms: two-stage and one-stage approaches. Two-stage object detection algorithms, or region-based detection algorithms, involve two phases: first, region proposal, where a series of candidate bounding boxes that may contain objects of interest are generated; second, classification and localization are performed using a convolutional neural network. Common algorithms include R-CNN[8], SPPNet[9], Fast R-CNN[10], Faster R-CNN[11], Mask R-CNN[12], and Cascade R-CNN[13].

Regression-based object detection algorithms operate in a single stage, directly regressing the predicted objects without needing a candidate region proposal phase. Instead, feature extraction, object classification, and location regression are all performed within a single convolutional network. This process, which calculates object positions and categories in one backward pass, is thus referred to as one-stage object detection. Compared

Institute of Big Data and Artificial Intelligence, Southwest Forestry University, Kunming, China. ✉email: zhliang@swfu.edu.cn

to two-stage object detection algorithms, one-stage algorithms are optimized for real-time detection, making them more suitable for tasks that require high-speed performance.

The YOLO[14] and the SSD[15] are exemplary regression-based object detection algorithms. Each has detection accuracy and speed advantages, catering to different application scenarios.

The YOLO algorithm is a highly efficient real-time object detection method widely used in computer vision. Its core advantage lies in its ability to transform object detection into a simplified regression problem, allowing the neural network model to directly and rapidly predict the categories and precise locations of all objects within an image. This innovation distinctly sets it apart from traditional methods that require generating candidate regions followed by classification—a more cumbersome process. By requiring only a forward pass, YOLO can simultaneously detect all objects in an image, significantly enhancing processing speed and responsiveness. Additionally, YOLO leverages global information from the entire image during prediction, offering specific advantages when detecting multiple objects and handling complex scenes.

Although YOLO boasts fast detection speed, its accuracy has traditionally been lower. In recent years, successive versions of YOLO have undergone continuous improvements, aiming to achieve higher accuracy and speed, thereby adapting to a broader range of application scenarios. YOLOv2[16] introduced enhancements such as Batch Normalization(BN) and anchor boxes, significantly improving accuracy and speed. By incorporating multi-scale prediction, YOLOv3[17], increased detection precision, particularly for small objects. YOLOv4[18] optimized training strategies, adding features like Mosaic data augmentation and the CSPDarknet53 backbone network, further boosting performance. YOLOv5 introduced strategies such as Mosaic data augmentation and adaptive anchor box calculation, improving the model's generalization ability while balancing accuracy and speed well. YOLOv6[19] adopted an anchor-free design and integrated the RepVGG[20] module. It offers performance and speed improvements over YOLOv5, especially on devices with limited computational power, making it well-suited for industrial-grade applications. YOLOv7[21] introduced the E-ELAN module and incorporated model re-parameterization into the network architecture, enhancing accuracy and generalization capabilities. YOLOv8 optimized the network architecture, making feature extraction and fusion more efficient, significantly boosting detection accuracy. Its highly efficient and lightweight design makes it suitable for object detection tasks across various scenarios.

Currently, most warehouse management practices involve monitoring and managing through manual review and playback of surveillance camera footage[22]. This approach is not only labor-intensive and resource-consuming but also significantly inefficient. To address these issues and promote the advancement of intelligent and informative warehouse management, this study proposes the use of object detection for staff and forklifts in food freezer warehouses. This detection method optimizes daily warehouse operations, addresses efficiency and intelligence challenges in warehouse management, and enhances the overall efficiency and quality of logistics operations.

Based on the above analysis, this study proposes a lightweight object detection model, YOLOv8-RSS, with the following key contributions:

(1) We introduce a novel module, C2f_RDB: A newly designed C2f_RDB module replaces the original C2f module in the YOLOv8 Backbone for feature extraction. This significantly enhances the model's feature extraction capabilities, improving detection accuracy while reducing parameter count, resulting in a more lightweight and efficient model without compromising performance.
(2) Integration of the SimAM Attention Mechanism: This mechanism allows the model to focus more effectively on critical information in the image, achieving a substantial improvement in detection accuracy without adding extra parameters.
(3) Adoption of Soft-NMS: The traditional NMS method is replaced with Soft-NMS, improving the model's detection performance in complex scenarios. This enhancement provides more reliable technical support for the intelligent management of food freezer warehouses.

The structure of this study is as follows: Sect. 2 provides an overview of the self-constructed dataset and details the proposed method. Section 3 discusses the evaluation strategy and presents the experimental results, followed by a comprehensive model evaluation. Finally, Sect. 4 offers concluding remarks and suggests potential directions for future research.

Materials and methods

The technical roadmap of this study is illustrated in Fig. 1, which details the complete process from image data collection to model evaluation. In the following sections, we will detail the specifics of each stage.

Dataset description

Currently, public datasets for warehouse scenarios are relatively scarce. This study selected a complex food freezer warehouses as the research background to meet research needs and created a food freezer warehouses dataset. The dataset was sourced from a food freezer warehouses in Yuxi, Yunnan Province, using data from three surveillance cameras positioned in different directions. The collected data, as shown in Fig. 2, consists of 4,463 images, each with a resolution of 1920×1080. This study has obtained informed consent from the warehouse manager, permitting the publication of images in open-access online publications.

After completing data collection and organization, the next step is image annotation. It is important to note that food freezer warehouses must maintain a constant low temperature of -40 °C to preserve the quality of stored goods. In this environment, workers are required to wear specialized antifreeze and thermal clothing when entering the warehouse. However, during our field research, we observed that some workers failed to

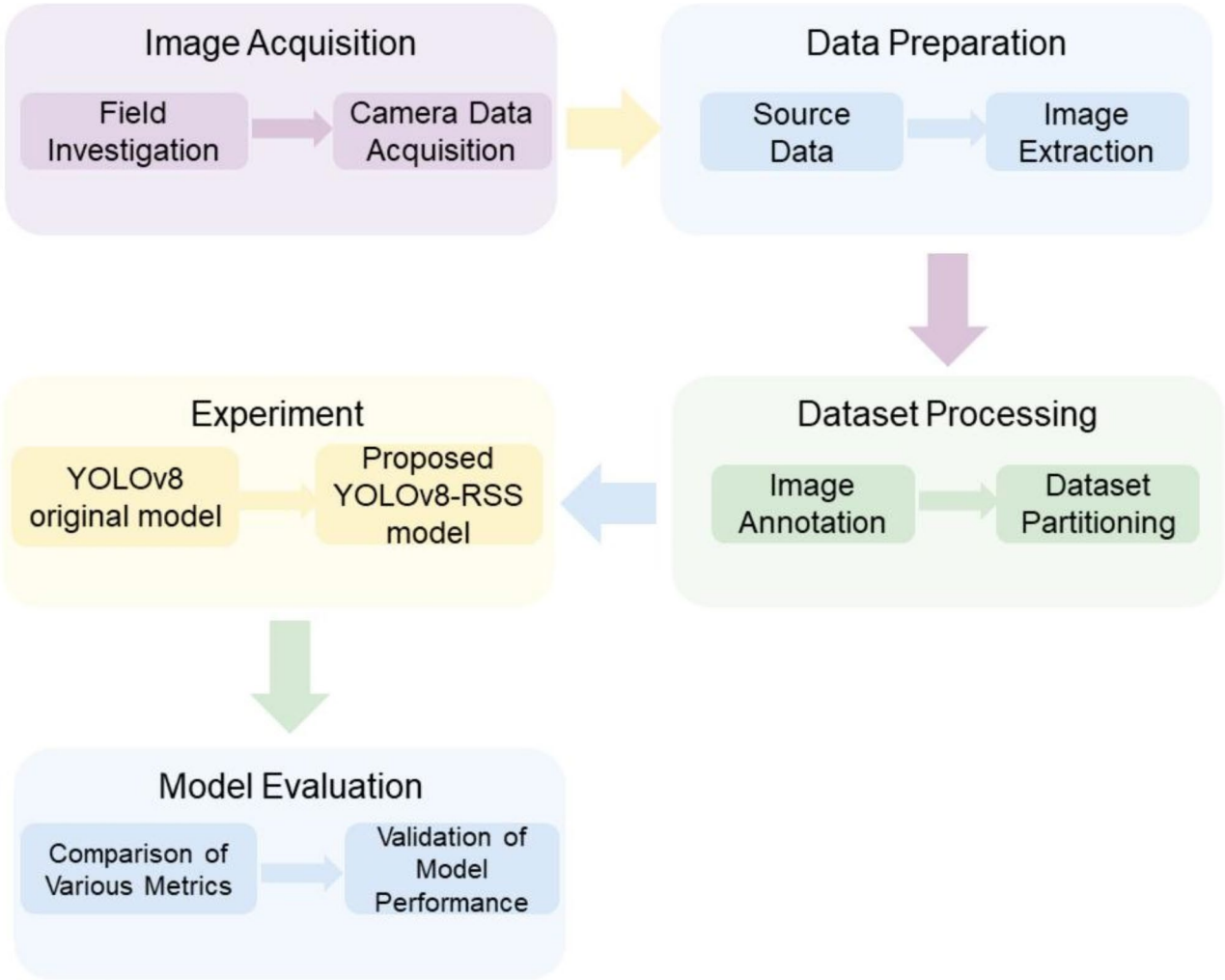

Fig. 1. technical roadmap.

adhere to these safety regulations, entering the freezing warehouse without the necessary protective gear, which poses a severe safety risk.

For the specific scenario of the food freezer warehouses and the research needs, five categories of objects were annotated in the data: work clothes, not work clothes, manual forklifts, electric push forklifts, and stainless steel forklifts. The five object categories in the dataset are shown in Fig. 3. After annotation, the dataset images were randomly divided into training, validation, and test sets in a 7:1:2 ratio, and the network model was trained accordingly.

Overview of YOLOv8 architecture

The YOLOv8 model, launched in 2023 by the Ultralytics team and other contributors, represents the next generation of real-time models. Compared to previous versions, YOLOv8 performs object detection, instance segmentation, and image classification more rapidly and expands its capabilities to include Pose estimation and Oriented Bounding Boxes Object Detection (OBB). This demonstrates a more comprehensive performance. The network model architecture of the YOLOv8 algorithm is illustrated in Fig. 4.

In YOLOv8, the backbone utilizes the CSPDarkNet architecture, primarily consisting of Conv, C2f, and SPPF modules. The Conv module, composed of Conv2d, Batch Normalization, and SiLU[23] functions as the standard convolutional unit in the YOLOv8 algorithm. An improvement on the C3 module, the C2f introduces additional parallel gradient flow branches, thereby capturing richer gradient flow information while preserving a lightweight structure. This design speeds up the model's convergence and enhances its overall performance while lowering the computational burden. The Spatial Pyramid Pooling Fast (SPPF) module, evolved from Spatial Pyramid Pooling (SPP), cleverly transforms feature maps of arbitrary sizes into fixed-size feature vectors. This design aims to achieve the perfect integration of features of different sizes, effectively addressing the limitation of convolutional neural networks having fixed input image size requirements. By doing so, we can capture and utilize more abundant information, thereby improving the model's performance.

Fig. 2. Data collected through 3 different surveillance cameras.

The primary function of the Neck in YOLOv8 is feature fusion, which efficiently combines feature maps from various layers. By leveraging the FPN[24] and PAN[25] structures, it enhances feature fusion through both top-down and bottom-up pathways. By performing feature fusion in the Neck, the YOLOv8 model can acquire richer and more accurate feature information, leading to higher accuracy in object detection tasks. This design exhibits greater robustness in scenes with complex backgrounds and multiple classes of objects, allowing YOLOv8 to handle various detection tasks and datasets flexibly.

The Head of YOLOv8 employs a Decoupled Head architecture, which separates classification and regression tasks, enhancing detection accuracy and flexibility. YOLOv8 has moved away from the anchor-based approach used in earlier versions, adopting an anchor-free strategy. Instead of relying on predefined anchors, the model directly predicts object center points, reducing the number of bounding boxes during prediction and speeding up the Non-Maximum Suppression (NMS) process. This shift enhances computational efficiency while maintaining high detection performance.

Presented methodology

To address the urgent need for intelligent warehouse management, this study proposes a new object detection model: YOLOv8-RSS. The architecture of the YOLOv8-RSS model is illustrated in Fig. 5. It enhances detection accuracy and achieves model lightweight, providing robust technical support for the intelligent management of modern warehouses.

First, we introduced a novel module, C2f_RDB, which replaces the C2f module in the YOLOv8 Backbone. This modification significantly enhances the model's feature extraction capability while reducing the number of parameters. Next, we integrated the SimAM[26] attention mechanism into the final layer of the Backbone, allowing the model to better focus on critical information in the image, thereby improving detection accuracy without increasing the parameter count. Finally, we replaced the traditional NMS method with Soft-NMS[27], demonstrating greater robustness in handling overlapping objects. This is particularly advantageous for detecting densely packed or occluded objects commonly encountered in food freezer warehouses.

C2f_RDB module

This study focuses on developing an accurate and lightweight model to balance precision and efficiency perfectly. Inspired by the EMO model[28] and UniRepLKNet[29], we have thoroughly re-evaluated and redesigned the C2f module in the YOLOv8 algorithm, successfully creating an innovative C2f_RDB module. This new module will replace the original C2f module in the Backbone, serving as a core feature extraction unit. This innovative design is expected to enhance the model's performance while maintaining its lightweight and efficient characteristics. The architecture of the C2f_RDB module is illustrated in Fig. 6.

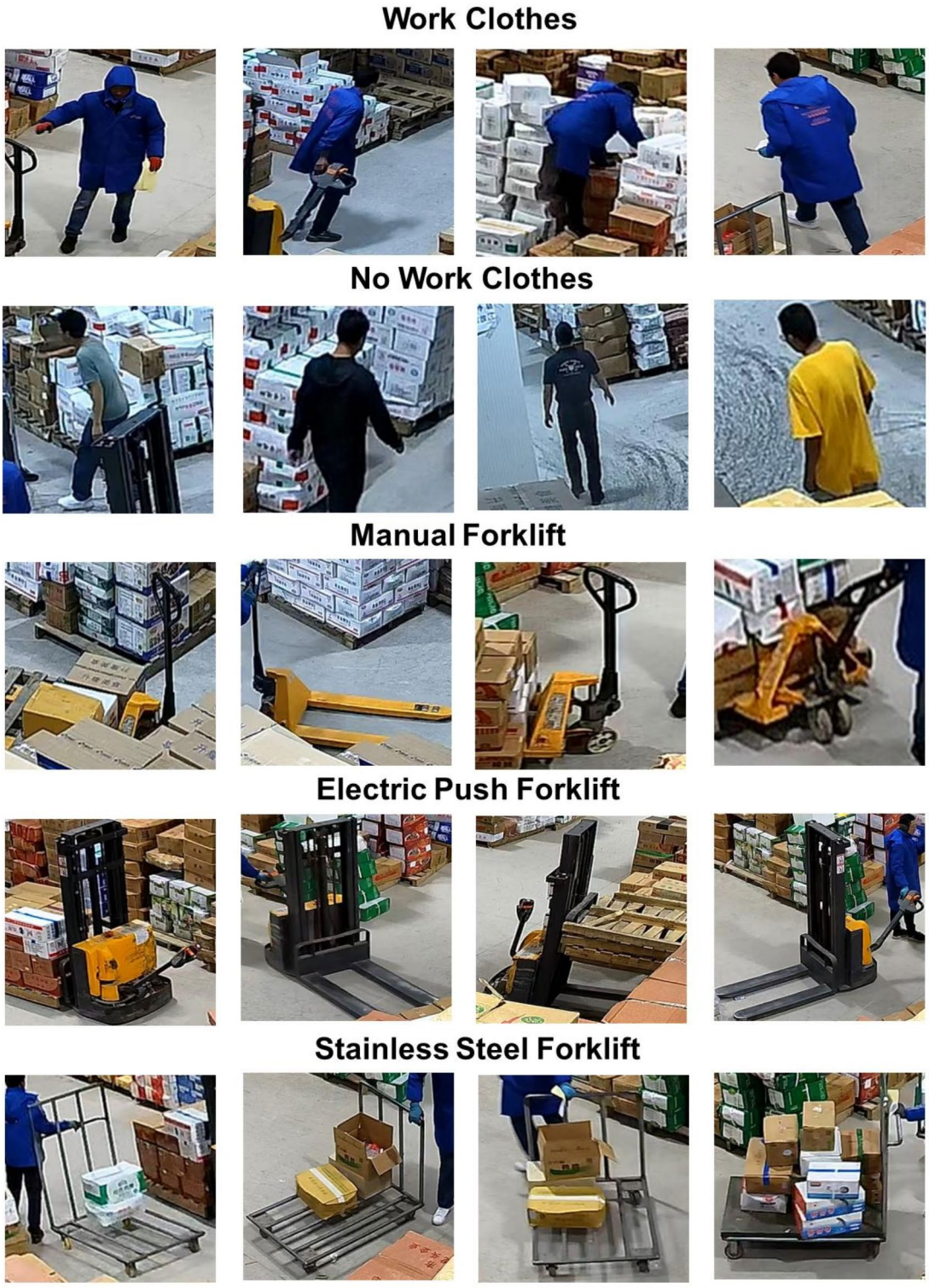

Fig. 3. The five categories of objects classified in the dataset.

The RDB module is an innovative improvement based on the Inverted Residual Mobile Block from the EMO model[28]. We introduced the DRB[29] large-kernel convolution into this structure, resulting in the new RDB module. The specific structure of the RDB module is illustrated in Fig. 7.

The RDB module, built on the inverted residual architecture, enhances information flow and captures long-range relationships while maintaining a lightweight structure. The multi-head self-attention mechanism, integrated into the RDB, further improves the model's ability to capture global dependencies across input data segments, enhancing its understanding of complex patterns. By combining self-attention with local convolution

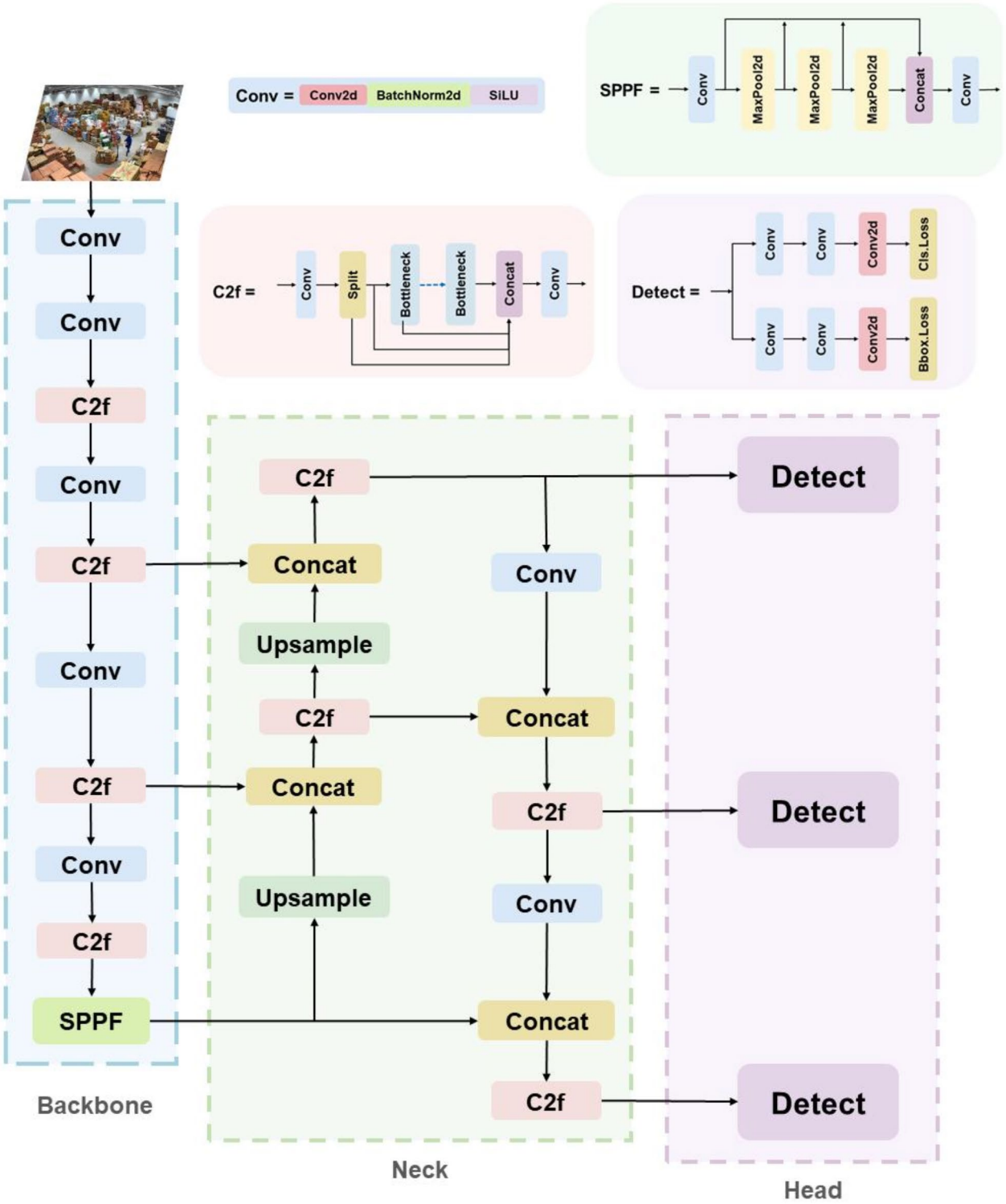

Fig. 4. YOLOv8 model network architecture and its specific components used to build the architecture.

operations and residual connections, the model gains more prosperous and dynamic feature representations, allowing it to process local and global information simultaneously. At the core of the multi-head self-attention mechanism is the computation of weighted representations using Queries (Q), Keys (K), and Values (V). First, the input x undergoes a linear transformation to obtain the Q, K, and V: $Q = W_Q x$, $K = W_K x$, $V = W_V x$, where W_Q, W_K, and W_V are the weight matrices for the linear transformations. First, we compute the scaled dot-product attention scores. To prevent the dot-product results from becoming too large due to increased dimensions, which could lead to gradient vanishing or explosion, we apply a scaling method. The dot-product

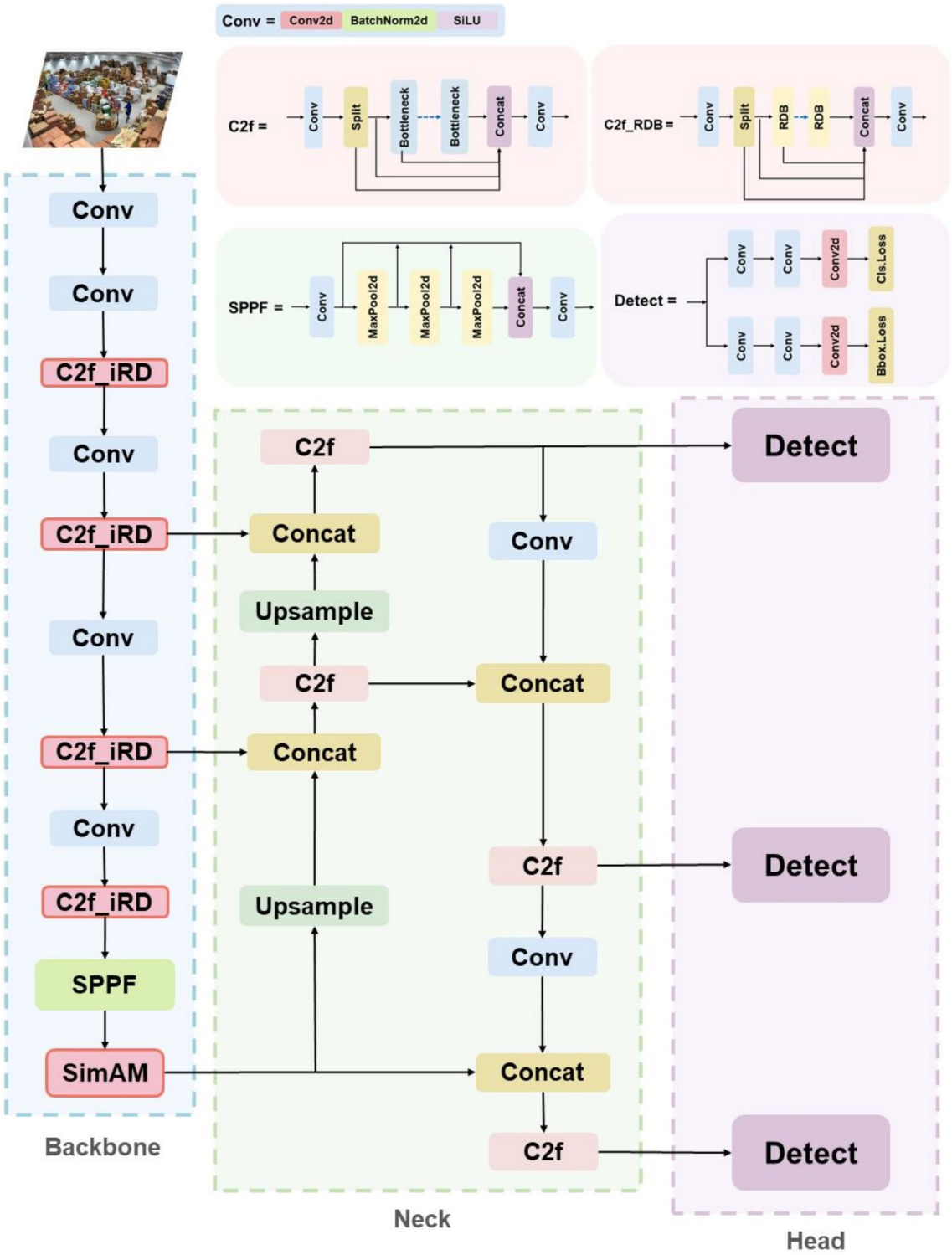

Fig. 5. YOLOv8-RSS model network architecture.

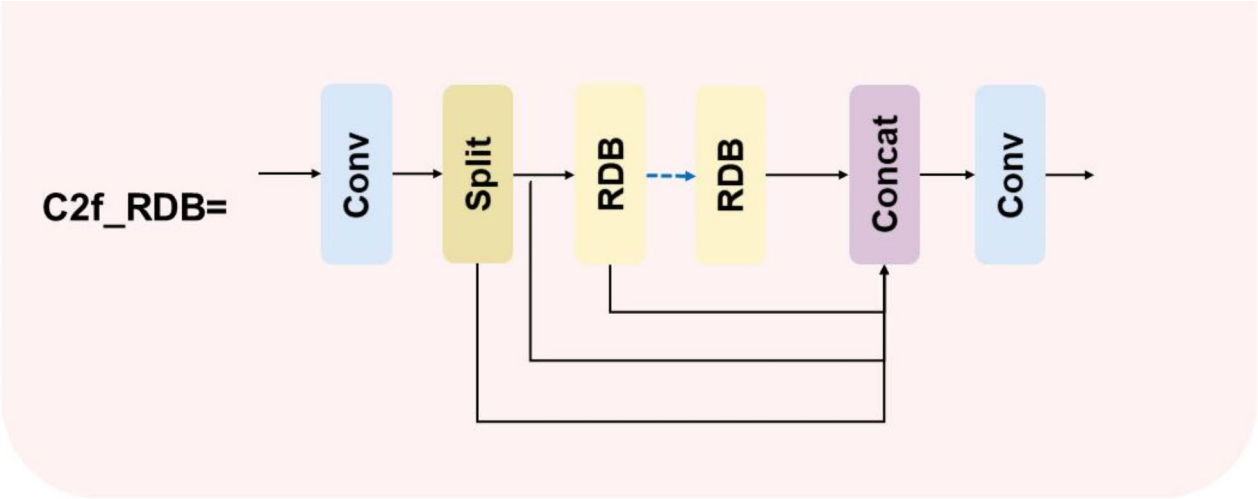

Fig. 6. C2f_RDB module architecture.

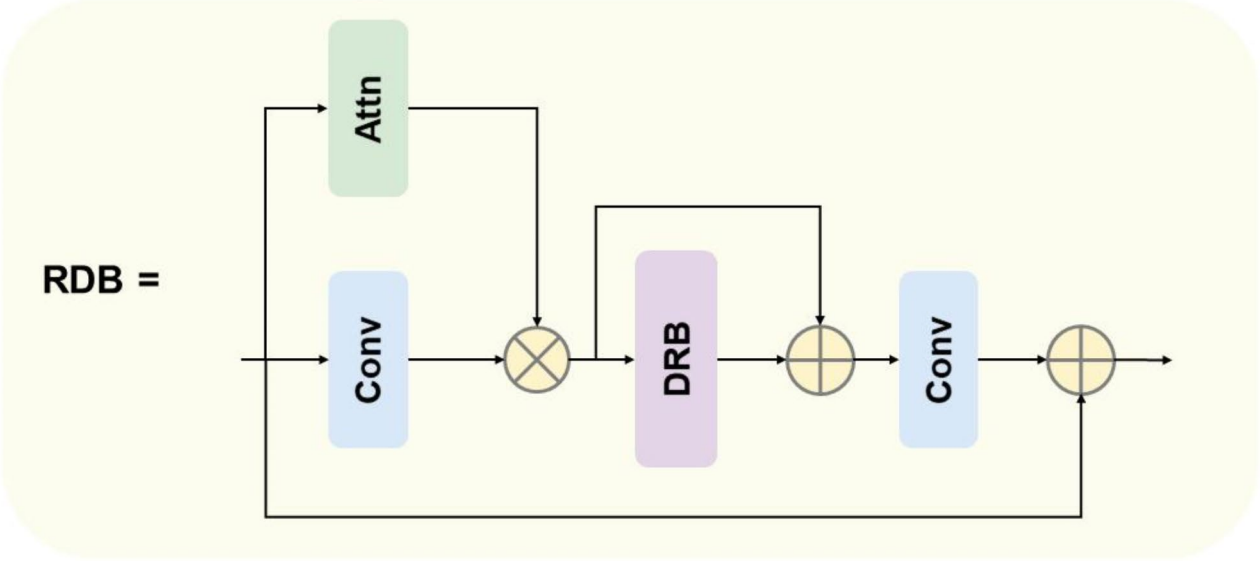

Fig. 7. RDB module architecture.

result is divided by the square root of the dimensionality d_{head} of the attention head. The specific formula is given in Eq. 1.

$$\text{Attention Scores} = \frac{\text{QK}^{\text{T}}}{\sqrt{\text{d}_{\text{head}}}} \tag{1}$$

This scaling operation allows the model to dynamically compute the similarity between input features and adaptively focus on critical parts of the input features. Additionally, the scaling enhances computational stability, effectively preventing numerical instability issues that could arise from high-dimensional spaces.

To get the attention weights, the attention scores are further normalized using the SoftMax algorithm. To enhance the regularization effect, Dropout is applied to these attention weights. Then, using these adjusted attention weights, a weighted sum of the values is performed to achieve a dynamic feature integration process. Combining the aforementioned steps, the complete attention mechanism formula is shown in Eq. 2.

$$Attn = Dropout\left(\text{softmax}\left(\frac{(W_Q x)(W_K x)^T}{\sqrt{d_{\text{head}}}}\right)\right) \times (W_V x) \tag{2}$$

In the DRB module, in addition to the core large-kernel convolution operation, parallel dilated convolution techniques are also cleverly introduced. Structural re-parameterization is employed during deployment, merging multiple convolution kernels into a single large-kernel convolution. The non-dilated large-kernel layers are improved by using dilated small-kernel convolution layers. Multiple dilated convolutions and BN layers (batch normalization) are used in training mode. In deployment mode, these convolutions and BN layers are fused into a more superficial convolution layer to improve inference efficiency. The structure of the DRB module is illustrated in Fig. 8.

Before performing the convolution operation, padding is applied to ensure that the input and output sizes remain consistent. The padding calculation formula is provided in Eq. 3:

$$\mathrm{p} = \frac{r \cdot (k-1)+1}{2} \tag{3}$$

Dilated convolution increases the receptive field by inserting holes between the convolutional kernel elements. The formula for dilated convolution is given in Eq. 4, where x is the input, w is the convolution kernel, r is the dilation rate, and k is the size of the convolution kernel.

$$Conv\,(x, w, r) = \sum_{k=1}^{K} x\,(i + r \cdot k) \cdot w_k \tag{4}$$

Batch normalization is used to stabilize and accelerate the training process. The batch normalization formula is given in Eq. 5, where μ is the batch mean, σ^2 is the batch variance, ε is a small constant to avoid division by zero, and γ and β are the scale and shift parameters of batch normalization, respectively.

$$BN\,(x) = \gamma \cdot \frac{x-\mu}{\sqrt{\sigma^2+\varepsilon}} + \beta \tag{5}$$

Finally, the convolution and batch normalization will be fused and the formula is shown in Eq. 6

$$y = \sum_{r,k} \left(\gamma \cdot \frac{w_k}{\sqrt{\sigma^2+\varepsilon}} \cdot x\,(i + r \cdot k + \mathrm{p})\right) + \left(\beta - \gamma \cdot \frac{\mu}{\sqrt{\sigma^2+\varepsilon}}\right) \tag{6}$$

The model's ability to extract features is enhanced by this design, which makes it more effective in capturing sparsely distributed features across space. It maintains model performance stability while significantly reducing computational load and memory usage during inference, bringing substantial convenience and advantages to practical applications.

The purpose of the C2f_RDB module design is to optimize the network's ability to capture image data features while reducing computational resource consumption. This ensures that the model maintains high accuracy and achieves higher computational efficiency. The structure of the C2f_RDB module, which combines inverted residuals and dilated convolutions, effectively enhances feature propagation and reuse. This allows the network to learn richer and deeper feature representations, improving the model's generalization ability and robustness.

SimAM attention mechanism

In this study, we include the SimAM[26] attention mechanism into the final layer of the Backbone to enhance the model's feature representation capabilities in complex environments and reduce noise interference from surrounding elements. SimAM design draws inspiration from the human brain's cooperative use of spatial and channel attention when processing visual information. Leveraging this principle, the SimAM mechanism evaluates each neuron's relative importance, generating three-dimensional attention weights from feature maps. Assessing the relevance of each neuron is a crucial task. In visual neuroscience, neurons that convey the most

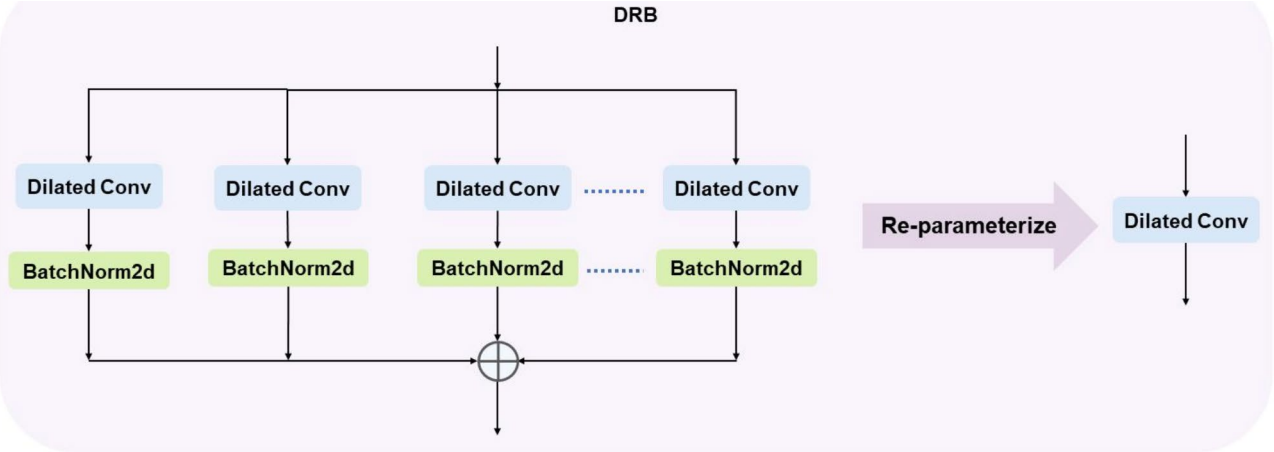

Fig. 8. DRB module architecture.

information often display distinct firing patterns and can inhibit the activity of neighboring neurons, a process known as spatial inhibition. The authors define the energy function for each neuron, as shown in Eq. 7.

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(w_t x_i + b_t))^2 + (1-(w_t t + b_t))^2 + \lambda w_t^2. \tag{7}$$

In the formula, t and x_i represent the target neuron and other neurons of the input feature X, where $\mathrm{X} \in \mathbb{R}^{C\times H\times W}$. The index i refers to the spatial dimension, and $M = H \times W$ denotes the number of neurons in a single channel. The w_t and b_t represent the weight and bias, respectively, during the transformation of a specific neuron.

By calculating the mean and variance of w_t, b_t and all neurons, the formula for minimal energy can be obtained, as shown in Eq. 8.

$$e_t^* = \frac{4\left(\widehat{\sigma}^2+\lambda\right)}{(t-\widehat{\mu})^2+2\widehat{\sigma}^2+2\lambda} \tag{8}$$

In Eq. 10, the $\widehat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$, $\widehat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i-\widehat{\mu})^2$. The final SimAM attention mechanism is optimized as shown in Eq. 9.

$$\widetilde{\mathrm{X}} = sigmoid\left(\frac{1}{\mathrm{E}}\right) \odot X \tag{9}$$

Soft-NMS

NMS is primarily used to clean up overlapping and redundant bounding boxes generated during object detection and to select the highest-scoring box for each object class, ensuring that only one bounding box is output for each target. This is an essential part of the object detection workflow. The process of traditional NMS is illustrated in Fig. 9.

In traditional NMS, an IOU threshold is first set when handling object detection tasks. Their scores then sort the bounding boxes output from the detection process in descending order. The highest-scoring bounding box

Fig. 9. NMS processing bounding box process.

is selected first. If other bounding boxes have an IOU value with this highest-scoring box that exceeds the preset threshold, their scores are directly set to zero, meaning they are suppressed.

While traditional NMS effectively reduces redundant and overlapping bounding boxes, it has limitations. When two objects of the same class are closely positioned, NMS often suppresses the bounding box with the lower score due to high overlap, leading to missed detections and a decline in detection accuracy.

In this study, the food freezer warehouse environment presents unique challenges, as workers or forklifts may overlap while performing tasks. In such cases, traditional NMS might overly suppress lower-confidence bounding boxes, resulting in missed detections of highly overlapping objects and reducing the algorithm's accuracy.

To address the issues above, we employ Soft-NMS[27] to replace the traditional NMS method. Soft-NMS handles occluded or highly overlapping targets by gradually decreasing their confidence scores based on the degree of overlap rather than simply deleting the bounding boxes with an IOU greater than the set threshold. This approach effectively retains partial information on occluded or overlapping targets, improving object detection accuracy without adding extra parameters.

The traditional NMS formula is shown in Eq. 10. Here, s_i represents the score of the current bounding box, b_i is the bounding box to be processed, M is the bounding box with the highest current score, and N_t is the IoU threshold. According to this formula, traditional NMS directly sets the scores of windows with an IoU greater than the threshold to zero.

$$s_i = \begin{cases} s_i, \text{iou}\,(M, b_i) & < N_t \\ 0, \text{iou}\,(M, b_i) & \geq N_t \end{cases} \tag{10}$$

In contrast, the Soft-NMS method suggests penalizing and attenuating the score s_i of the bounding box. The linear weighting formula is shown in Eq. 11.

$$s_i = \begin{cases} s_i, \text{iou}(M, b_i) & < N_t \\ s_i(1 - \text{iou}(M, b_i)), \text{iou}(M, b_i) & \geq N_t \end{cases} \tag{11}$$

The larger the IoU between b_i and M, the more s_i decreases. However, this approach is discontinuous regarding overlap; the penalized attenuation function can easily disrupt the score ranking order. A reasonable penalization function should gradually transition, with higher penalties for high overlap and lower penalties for low overlap. Considering the above situation, Soft-NMS paper proposed a new Gaussian penalty function as shown in Eq. 12.

$$s_i = s_i e^{-\frac{\text{iou}(M, b_i)^2}{\sigma}}, \forall\, b_i \notin D \tag{12}$$

Results and discussion

Performance evaluation metrics

This study proposes a high-precision yet lightweight model, evaluated primarily through four key metrics: parameter, FLOPs, mAP@0.5, and mAP@0.5:0.95.

The parameter directly affects the model's storage requirements, computational efficiency, and training time. Therefore, when designing and selecting an object detection model, the parameter is crucial.

In object detection, FLOPs (Floating Point Operations Per second) are employed to compare the computational efficiency of different models. A lower FLOPs value indicates that the model can perform computations faster under the same hardware conditions, making it suitable for use in resource-constrained environments.

A key performance indicator for object detection model evaluation is mAP (mean Average Precision). Each category's AP (Average Precision) values are averaged to determine it. Precision (P) and Recall (R) values are derived from the confusion matrix, which contains four key indicators TP (True Positive), FN (False Negative), FP (False Positive), TN (True Negative).

The proportion of true positive samples among all samples the model has determined to be positive is known as precision (P). The formula is given by Eq. 13.

$$P = \frac{TP}{TP + FP} \tag{13}$$

The proportion of true positive samples the model correctly detects relative to the total number of positive samples is known as recall (R). The formula is given by Eq. 14.

$$R = \frac{TP}{TP + FN} \tag{14}$$

The AP value is obtained by plotting each category's PR (Precision-Recall) curve and calculating the area under the curve. The formula for AP is shown in Eq. 15.

$$AP = \int_0^1 P\,(R)\, dR \tag{15}$$

The metrics mAP@0.5 and mAP@0.5:0.95 are specific instances of mAP that evaluate model performance under different Intersections over Union (IoU) thresholds. mAP@0.5 measures detection accuracy with a lower IoU threshold of 0.5, which allows for less stringent localization and emphasizes the model's ability to identify objects with some overlap with the ground truth boxes. In contrast, mAP@0.5:0.95 provides a more comprehensive

evaluation of detection performance by considering a range of IoU thresholds, offering a broader assessment of the model's accuracy across varying degrees of overlap. The formula for mAP is shown in Eq. 16.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{16}$$

Experimental results

Ablation study

In deep learning-based object detection, ablation studies are a crucial experimental method primarily used to evaluate the model and the contributions of newly added modules to the overall performance. Through this approach, we can gain deeper insights into the model and validate the effectiveness of various modules. In this study, we designed a total of eight ablation experiments. The batch size was set to 32, the optimizer used was Adam, and the number of epochs was set to 300. Table 1 presents the results of ablation experiments conducted on the food freezer warehouse dataset, with the bolded entries representing the baseline model data and the results of our proposed model.

In this study, YOLOv8n was used as the baseline for comparative analysis with other experimental results. First, after incorporating the C2f_RDB module, the model's parameters decreased by 0.05 M, FLOPs reduced by 0.8G, mAP@0.5 increased by 0.4%, and mAP@0.5:0.95 improved by 0.2%. Second, when SimAM and Soft-NMS were added separately, the model's parameter and FLOPs remained unchanged compared to the baseline model, but the accuracy improved.

Ultimately, the optimal experimental results were achieved with our proposed model, which utilized all three modules. Compared to the baseline model, this model reduced parameter counts by 0.05 M, decreased FLOPs by 0.8G, and improved mAP@0.5 by 1.4% and mAP@0.5:0.95 by 3.9%. This model is more lightweight and significantly enhances accuracy, fully meeting the requirements of the f food freezer warehouses.

The summarized PR results of mAP@0.5 for each part of the ablation study are shown in Fig. 10. The figure demonstrates the superior performance of our model.

Comparative experiments

We conducted comparison tests on the food freezer warehouses dataset against other advanced models to confirm the efficacy of our model. The experimental results are presented in Table 2. In these comparative experiments, we primarily selected other shallow network models from the YOLO algorithm series and other lightweight models from the object detection domain for comparison with our proposed model. The results show that our model outperforms similar models in achieving a lightweight design with excellent detection accuracy.

Visualization of results

A graphic comparison of the detection outcomes between our model and the YOLOv8n model is shown in Fig. 11 (for better display, some images have been cropped). Specifically, the second and third columns compare the detection accuracy of the two models, with different colors representing different categories. The fourth and fifth columns mainly compare the results of TP, FP, and FN between the two models. In these columns, green bounding boxes indicate correctly detected results (TP), blue bounding boxes indicate false detections where the model detected an object but classified it incorrectly (FP), and red bounding boxes indicate missed detections where the model failed to detect an actual object (FN).

The fourth column primarily shows the detection results of the YOLOv8n model, where figures (a), (b), and (c) exhibit FP, and figures (d) and (e) show instances of FN. The fifth column displays the results of our model, clearly demonstrating a reduction in missed and false detections. This further proves the effectiveness and robustness of our model.

Heat map analysis

In object detection, heatmaps are a crucial visualization tool that clearly illustrates which areas of an image the model focuses on during object detection. This deepens our understanding of the model's behavior during detection, thereby optimizing model performance. This study used the heatmap visualization tool Grad-CAM[30] to visualize three different models. The visualization results are shown in Fig. 12.

Model							
YOLOv8n	C2f_RDB	SimAM	Soft-NMS	Params (M)	FLOPs (G)	mAP@0.5	mAP@0.5: 0.95
√				**3.0**	**8.1**	**0.914**	**0.693**
√	√			2.65	7.3	0.918	0.695
√		√		3.0	8.1	0.916	0.700
√			√	3.0	8.1	0.923	0.720
√	√	√		2.65	7.3	0.920	0.686
√	√		√	2.65	7.3	0.919	0.723
√		√	√	3.0	8.1	0.919	0.726
√	√	√	√	**2.65**	**7.3**	**0.928**	**0.732**

Table 1. Results of ablation experiments.

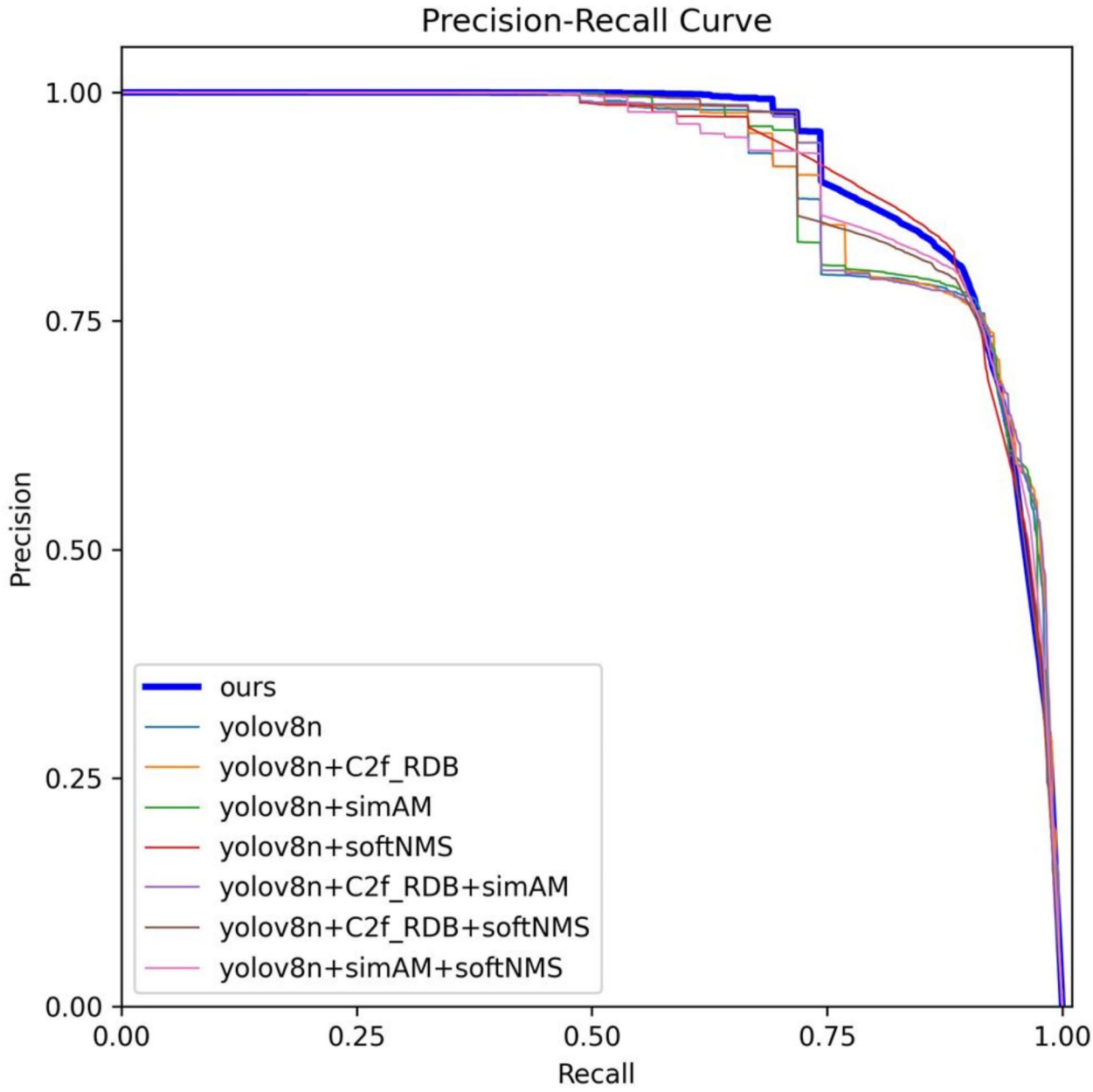

Fig. 10. Summary of PR at mAP@0.5 for each component in the ablation experiment.

Model	Epochs	Params	FLOPs	mAP@0.5	mAP@0.5: 0.95
YOLOv5n	300	2.50	7.2	0.914	0.685
YOLOv6n	300	4.23	11.8	0.915	0.696
YOLOv7-Tiny	300	6.02	13.2	0.915	0.647
RTMDet-Tiny	100	4.87	8.0	0.895	0.669
YOLOX-Tiny	100	5.03	3.2	0.841	0.523
YOLOv8n	300	3.0	8.1	0.914	0.693
Ours	300	2.65	7.3	0.928	0.732

Table 2. Comparative experimental results of the algorithms on the food freezer warehouses dataset.

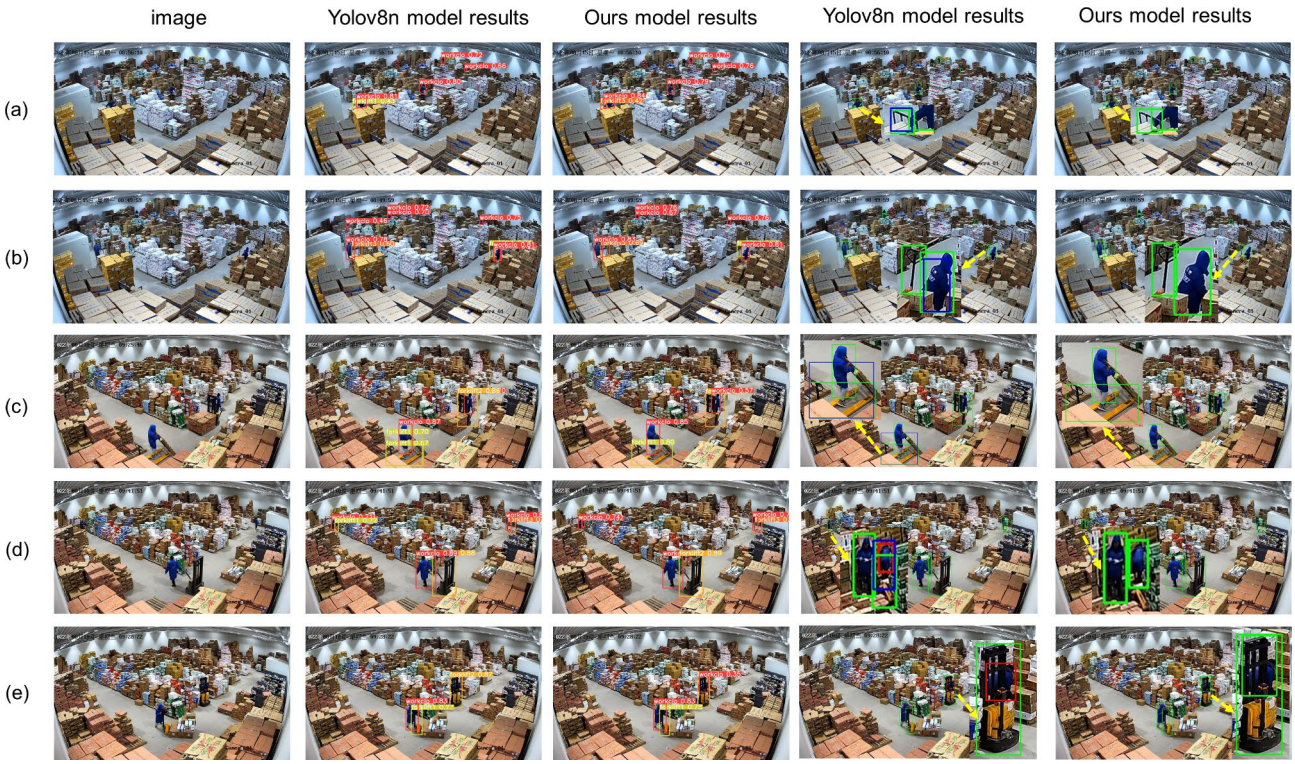

Fig. 11. Comparison between YOLOv8n model and our model in visualization of detection results.

From the visualization results, we can observe significant differences in the focus areas of the model when detecting different categories. When detecting humans, the model primarily focuses on the entire person. In contrast, when detecting forklifts, the model concentrates more on the forklift's forks. This finding provides important insights and assistance for further advancing the intelligent processes in food freezer warehouses.

Picture object counting

In the context of food freezer warehouses, this paper employs the YOLOv8-RSS model for object detection and object counting on images rapidly. The object counting results are demonstrated in Fig. 13, clearly showing the model's success in identifying and counting personnel and forklifts. The application of this technology significantly enhances our real-time monitoring capabilities of dynamic resources within the warehouse, enabling us to track operational status more accurately. This provides strong support for optimizing warehouse management and improving logistics efficiency. Moreover, it helps promptly identify personnel entering the freezer without work clothes, ensuring worker safety, preventing risks, and proactively avoiding safety incidents. Overall, it holds substantial practical significance and application value across multiple aspects.

Conclusion and future work

For the specific context of food freezer warehouses, the proposed YOLOv8-RSS model delivers significant improvements by detecting two core entities: personnel and forklifts. First, the model efficiently and accurately identifies their locations within the warehouse, enabling real-time operations monitoring, optimizing resource allocation, and improving operational efficiency. Second, its lightweight design ensures reliable real-time detection even in resource-constrained environments, providing practical technical support for intelligent warehouse management. Additionally, the model's detection capabilities help identify potential safety risks, such as personnel entering the freezer without proper work attire, thereby reducing the likelihood of safety incidents.

To validate the effectiveness of the YOLOv8-RSS model, we conducted experiments on a self-constructed food freezer warehouse dataset. Compared to the baseline model YOLOv8n, our model reduced the parameter count by 0.05 M, decreased FLOPs by 0.8G, increased mAP@0.5 by 1.4%, and improved mAP@0.5:0.95 by 3.9%. These experimental results powerfully demonstrate the effectiveness of our proposed model.

In this study, we applied the YOLOv8-RSS model to the practical task of object counting in images. Looking ahead, this model is expected to be deployed in actual food freezer warehouse scenarios to achieve object counting in images or videos according to the specific needs of management personnel. Thanks to the lightweight characteristics of the YOLOv8-RSS model, it can achieve real-time tracking during monitoring and is easy to deploy. By analyzing real-time data, warehouse managers can promptly adjust personnel allocation and forklift dispatch strategies, reducing wait times and minimizing resource idling, which, in turn, effectively enhances operational efficiency and accelerates logistics response times within the warehouse.

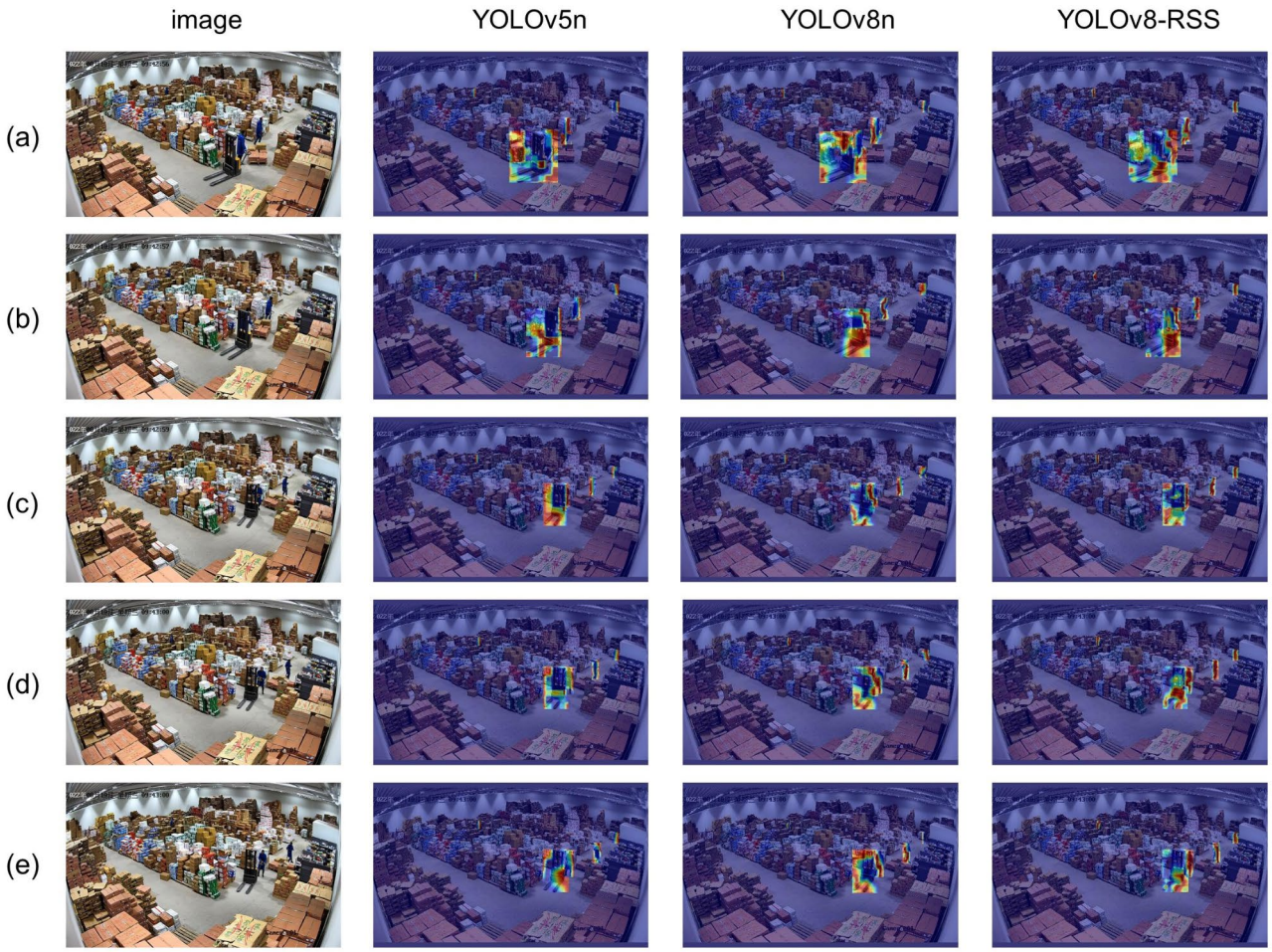

Fig. 12. Visualization results of heat maps with different models.

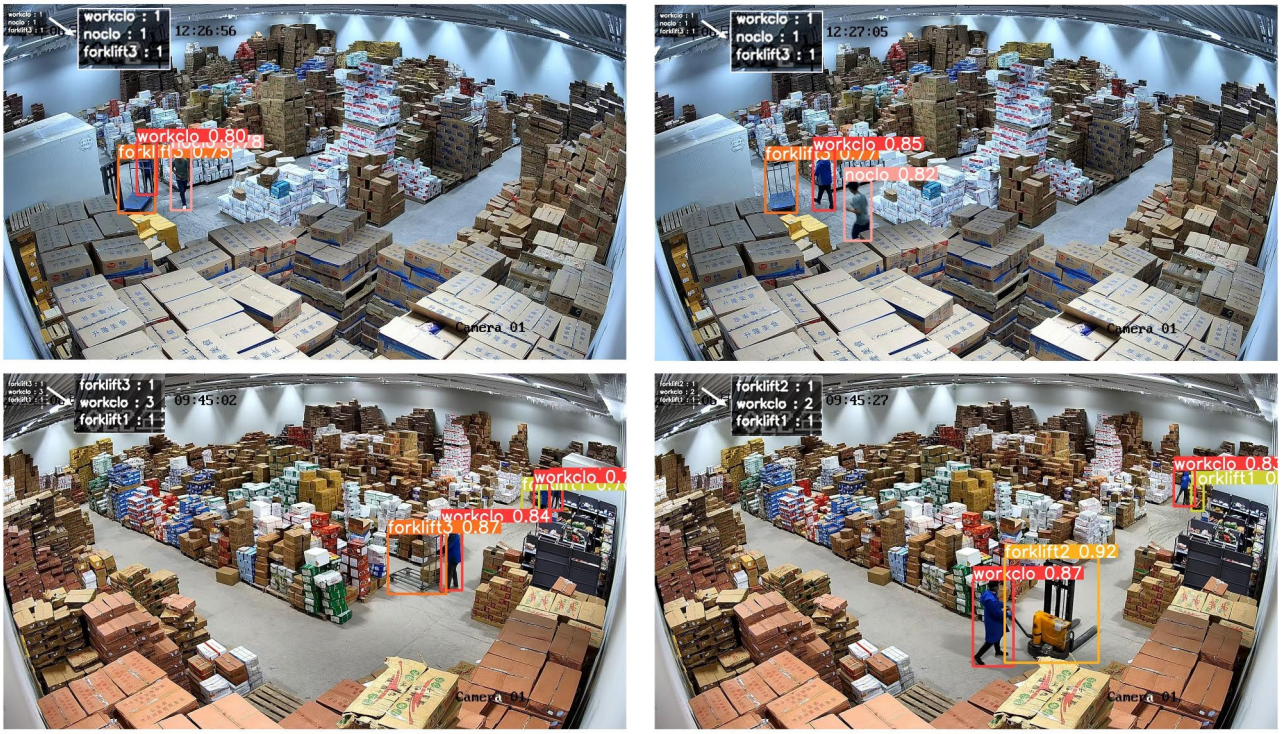

Fig. 13. YOLOv8-RSS model to implement object counting in pictures.

Data availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Received: 5 September 2024; Accepted: 13 January 2025
Published online: 17 January 2025

References

1. Kamali, A. Smart warehouse vs. traditional warehouse. *CiiT Int. J. Autom. Auton. Syst.* **11**(1), 9–16 (2019).
2. Zhen, L. & Li, H. A literature review of smart warehouse operations management. *Front. Eng. Manag.* **9**(1), 31–55 (2022).
3. Žunić, E., Delalić, S., Hodžić, K., Beširević, A. & Hindija, H. in *2018 14th Symposium on Neural Networks and Applications (NEUREL)*. 1–5 (IEEE).
4. Zou, Z., Chen, K., Shi, Z., Guo, Y. & Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **111**(3), 257–276 (2023).
5. Ye, M. et al. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(6), 2872–2893 (2021).
6. Chen, W., Huang, H., Peng, S., Zhou, C. & Zhang, C. YOLO-face: a real-time face detector. *Vis. Comput.* **37**, 805–813 (2021).
7. Huang, X. *et al.* in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 954–960.
8. Girshick, R., Donahue, J., Darrell, T. and Malik, J., Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587). https://doi.org/10.1109/CVPR.2014.81 (2014).
9. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1904–1916 (2015).
10. Girshick, R. in *2015 IEEE International Conference on Computer Vision (ICCV)*. 1440–1448 (2015).
11. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031 (2017).
12. He, K., Gkioxari, G., Dollár, P. & Girshick, R. in *2017 IEEE International Conference on Computer Vision (ICCV)*. 2980–2988.
13. Cai, Z. & Vasconcelos, N. J. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 1483–1498 (2019).
14. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 779–788.
15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C. & Reed, S. E. SSD: Single Shot MultiBox Detector. 21–37 (2016).
16. Redmon, J. & Farhadi, A. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6517–6525.
17. Redmon, J. & Farhadi, A. YOLOv3: An Incremental Improvement. (2018).
18. Bochkovskiy, A., Wang, C. Y. & Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. (2020).
19. Li, C. *et al.* YOLOv6: A single-stage object detection framework for industrial applications. (2022).
20. Ding, X. *et al.* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13733–13742.
21. Wang, C. Y., Bochkovskiy, A. & Liao, H. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. (2022).
22. Cheng, Y. *Abnormal Behavior Recognition of Logistics Warehousing Personnel Based on Computer Vision*, Anhui University of Science and Technology, (2021).
23. Elfwing, S., Uchibe, E. & Doya, K. J. N. n. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. **107**, 3–11 (2018).

24. Lin, T.-Y. *et al.* in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.
25. Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8759–8768.
26. Yang, L., Zhang, R.-Y., Li, L. & Xie, X. in *International conference on machine learning*. 11863–11874 (PMLR).
27. Bodla, N., Singh, B., Chellappa, R. & Davis, L. S. in *Proceedings of the IEEE international conference on computer vision*. 5561–5569.
28. Zhang, J. *et al.* in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 1389–1400 (IEEE Computer Society).
29. Ding, X. *et al.* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5513–5524 (2024).
30. Selvaraju, R. R. *et al.* in *Proceedings of the IEEE international conference on computer vision*. 618–626.

Acknowledgements

This study was supported by the Central Government-Guided Local Science and Technology Development Fund Project, "China-Myanmar Cross-Border Logistics and Trade Integration Service Platform" (Project No. 202307AB110009), and the Yunnan Province Major Science and Technology Project "Research and application demonstration of key blockchain technologies serving key industries" (Project No. 202002AD080002).

Author contributions

J.Y.: Manuscript writing, conceptualization, methodology, software, data collection. Z.L.: Conceptualization, Supervision, Funding acquisition. M.Q. and X.T.: Resources, Visualization, Software. F.X. and A.H.: Data curation, Investigation. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.L.