APEM journal
Advances in Production Engineering & Management
Volume 21 | Number 2 | June 2026 | pp 163–177
https://doi.org/10.14743/apem2026.2.567
ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Industrial multi-object detection for automotive battery manufacturing using a CBAM-enhanced YOLOv5 framework

Li, Y.X.[a,*], **Chen, H.**[a], **Zhou, L.**[b]

[a]School of Intelligence Technology, Geely University of China, Chengdu, P.R. China
[b]Huayou Cobalt Co., LTD, Chengdu, P.R. China

ABSTRACT

The mixing stage of automotive battery production requires reliable monitoring of raw material types, personnel actions, and correct tool use. However, accurate multi-scale object detection in complex scenes remains challenging because of background interference and the requirements of embedded deployment and real-time operation. This study proposes an enhanced YOLOv5 framework for industrial multi-object detection. The method adopts a dual-stage feature-enhancement strategy designed to improve robustness while limiting parameter count and computational complexity. First, the Convolutional Block Attention Module (CBAM) is embedded in the Backbone and Neck of YOLOv5 to provide multi-granularity feature enhancement and improve the detection of small objects, such as tools. Second, the conventional CIOU loss is replaced with the Focal-EIOU loss function to optimize bounding-box regression, reduce false detections across multiple target scales, and accelerate model convergence. Finally, K-means clustering is applied to target geometric features to generate specialized anchor-box parameters better suited to industrial scenarios. Experimental results on an automotive battery production-site dataset show that the improved model's mAP@0.5 increased by 2.4 % compared with the original model, reaching 95.2 %, while mAP@0.5:0.95 improved by 3.6 %, reaching 78.5 %. The proposed framework provides a lightweight and reliable solution for multi-object detection in complex industrial environments and supports the development of intelligent visual monitoring systems for smart manufacturing.

ARTICLE INFO

Keywords:
Automotive battery manufacturing;
Industrial computer vision;
Multi-object detection;
Real-time object detection;
Deep learning;
YOLOv5;
Convolutional Block Attention Module (CBAM);
Dual-stage Feature Enhancement;
Focal-EIOU loss

**Corresponding author:*
liyunxue@bgu.edu.cn
(Li, Y.X.)

Article history:
Received 26 June 2025
Revised 22 June 2026
Accepted 25 June 2026

Content from this work may be used under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

1. Introduction

With the advancement of intelligent manufacturing, collaboration between humans and embedded intelligent agents is becoming increasingly close [1]. Artificial intelligence is being progressively introduced into production scenarios such as quality inspection, equipment monitoring, and process standardization [2, 3]. As core components of electric vehicles and energy storage systems, batteries must be manufactured to high-quality standards because their production quality directly affects product safety and market competitiveness. In complex production scenarios, comprehensive monitoring of personnel and the environment requires the recognition of multi-scale and small targets. During the mixing process in battery manufacturing, identifying raw material types, verifying that workers wear safety helmets, and monitoring the correct use of tools are important for ensuring process consistency. Traditional sensors have inherent limitations, while manual visual inspection is inefficient, subjective, and prone to fatigue. Object detection in

such complex scenarios therefore requires lightweight models that support embedded AI deployment and real-time factory operation [4].

Object detection algorithms have evolved from multi-stage detectors such as Faster R-CNN [5] to single-stage detectors such as SSD [6] and the YOLO series [7], progressively seeking a balance between accuracy and speed. Faster R-CNN achieves high-precision detection by generating candidate boxes through a Region Proposal Network (RPN) combined with Region of Interest Pooling (RoI Pooling). However, its complex two-stage process results in an inference speed that may not meet industrial real-time requirements. SSD directly predicts bounding boxes from multi-scale feature maps, significantly increasing detection speed. However, it is less sensitive to small objects, which can lead to missed detections of tiny targets such as defects [8]. In neural-network optimization, techniques such as L2 regularization and neuron dropout have been employed to prevent overfitting [9]. For feature extraction, common improvement approaches include wide convolution extraction, customized gating modules, and multi-layer progressive extraction [10].

Against this backdrop, the YOLO series has emerged as a mainstream choice for industrial applications because its end-to-end single-stage architecture and efficient network design provide a strong balance between accuracy and speed in product inspection, quality control, and defect detection [11-14]. Through lightweight strategies such as depthwise separable convolutions and Cross Stage Partial (CSP) connections, YOLOv5 significantly reduces model size. While maintaining high Mean Average Precision (mAP) [15, 16], it achieves an inference speed of up to 110 Frames Per Second (FPS), providing a feasible basis for deployment on edge devices. However, the original YOLOv5 still has several limitations in complex industrial scenarios. Background interference caused by metal reflections and motion blur from mechanical vibrations can reduce the robustness of feature extraction. In addition, the lower resolution of feature maps in deeper network layers can lead to a loss of detailed information for small objects. Conventional IoU-based loss functions may also model overlapping bounding boxes imprecisely, leading to false detections.

Attention mechanisms have recently played an important role in identifying and adaptively weighting relevant features across image-analysis and intelligent-manufacturing applications [17-19]. These include Squeeze-and-Excitation Networks (SE-Net) for extracting channel-wise feature importance, Efficient Channel Attention Networks (ECA-Net) for reducing parameter count and computational load [20, 21], and Coordinate Attention Networks (CA-Net) for lightweight mobile networks [22, 23]. Among these approaches, the Convolutional Block Attention Module (CBAM) cascades channel and spatial attention to adaptively adjust feature-map weights [24]. This enhances responses in target regions while suppressing background noise. More broadly, attention-based mechanisms have demonstrated their effectiveness in manufacturing and other complex application scenarios [25, 26].

With the emergence of Vision Transformer (ViT) applications [27], researchers have attempted to replace YOLO's CNN modules with Transformers. By establishing long-range dependencies through global self-attention, Transformers overcome the limited local receptive fields of convolution operations. More recent YOLO variants increasingly incorporate attention-based and Transformer-inspired modules to enhance feature extraction and multi-scale feature fusion. For instance, an improved YOLOv8 framework for pouch-battery defect detection incorporates a BiFormer attention mechanism into the feature-fusion stage to enhance content-aware feature representation [28]. However, the high computational complexity of Transformers conflicts with the real-time requirements of industrial applications, requiring a balance through structural pruning or hybrid architecture design.

The design of the loss function directly affects bounding-box regression accuracy and model convergence efficiency. Complete IOU (CIOU) loss improves on traditional IOU by accounting for bounding-box alignment through center-point distance and aspect-ratio penalty terms [29]. However, class imbalance remains a common challenge in industrial datasets [30]. Focal Efficient IOU (Focal-EIOU) Loss combines the dynamic weighting strategy of Focal Loss with the geometric awareness of EIOU [31]. By reducing the gradient contribution of easy samples and increasing the learning weight of hard samples, it improves the localization accuracy of small and densely distributed targets.

To address the requirements of multi-object detection in automotive battery production, this paper proposes an improved YOLOv5 framework based on the CBAM attention mechanism and Focal-EIOU Loss. The main improvements are summarised as follows:

- Dual-stage Feature Enhancement Network: First, a CBAM attention module is embedded in the Backbone of YOLOv5s, where the channel-attention branch captures multi-scale contextual information. Second, an attention mechanism is incorporated into the feature-fusion stage of the Neck network, where the spatial-attention branch improves small-object detection. This lightweight design effectively reduces computational complexity and parameter count.
- Focal-EIOU Loss Replacement: The original CIOU Loss is replaced with Focal-EIOU Loss. By introducing a dynamic focusing factor and an improved aspect-ratio penalty term, this modification reduces false detections of multi-scale targets while accelerating model convergence.
- Optimized Prior Anchor Boxes: The K-means algorithm is used to cluster the geometric features of targets and generate specialized prior anchor-box parameters that closely match industrial scenarios. This addresses the mismatch between the preset anchor boxes of the general COCO dataset and the actual target distribution, thereby improving localization accuracy.

Experiments on an automotive battery production dataset demonstrate that the proposed model substantially outperforms the original YOLOv5s in both background suppression and small-object detection accuracy. Furthermore, its lightweight architecture provides a deployable solution for multi-object detection in complex industrial environments.

2. Related work

Since YOLOv5 was released as an open-source project, its versions have been continuously updated. YOLOv5 is a single-stage object detection model available in four variants based on network depth and performance: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Among these, YOLOv5x has the largest number of parameters and achieves the highest detection accuracy, but it also has the slowest detection speed. Conversely, YOLOv5s has the fewest parameters and the smallest architecture, resulting in faster detection. Therefore, YOLOv5s is well suited to applications with stringent real-time performance requirements.

The YOLOv5s model primarily consists of three main components: Backbone, Neck, and Head. The input stage uses Mosaic data augmentation, which randomly scales, crops, and arranges images to enrich the detection dataset and reduce GPU usage. The YOLOv5s Backbone is mainly composed of CBS, CSP, and SPPF modules, whose primary function is to extract and fuse feature maps. The Neck uses FPN (Feature Pyramid Network) and PAN (Path Aggregation Network) structures to enhance feature fusion by combining information extracted from the Backbone. The FPN performs top-down upsampling so that lower-level feature maps contain stronger high-level semantic information, whereas the PAN performs bottom-up downsampling so that higher-level features retain positional information. These features are then fused, allowing feature maps at different scales to contain both semantic and positional information and thereby supporting accurate predictions for objects of varying sizes. The Head performs multi-scale object detection on the extracted feature maps and predicts bounding boxes for different object scales through multiple branches.

Based on the foregoing analysis, YOLOv5s has the fewest parameters, the highest inference speed, and the lowest computational complexity among the Transformer-encoder-based YOLO variants and other members of the YOLOv5 family considered here. In the complex battery-production scenario, it has 7 MB parameters, an activated memory footprint of 15 MB, a peak memory consumption of 20 MB, and a computational complexity of 8 GFLOPs (Giga Floating-Point Operations). Given its demonstrated performance in embedded deployments across multiple industrial settings, YOLOv5s is a suitable choice for multi-object detection in the battery-production mixing process. The multi-target classification results of YOLOv5s from different camera viewpoints are presented in Fig. 1.

Although YOLOv5s can detect multiple targets from different camera perspectives in battery-production scenes, missed detections still occur, and the recognition accuracy for small targets requires improvement. This limitation can be addressed by incorporating an attention mechanism to strengthen feature extraction.

Fig. 1 Multi-target classification by YOLOv5s under different camera viewpoints

3. Improved YOLOv5s method

3.1 Improved network modules

The modified components of the algorithm are shown in light green in Fig. 2. The input image is first processed by the YOLOv5s preprocessing module and resized to 640 × 640 pixels. The model then uses target anchor-box sizes clustered from data collected at the mixing production site, enabling the anchors to better match objects such as scissors and safety helmets.

To address the detection requirements of multi-scale and small objects, dual-branch CBAM attention modules are introduced to enhance the model's representation of spatial, multi-scale, and small-object features. To maintain a lightweight design, CBAM is inserted only in the deep feature layers of the YOLOv5s Backbone and at key fusion nodes of the Path Aggregation Network (PAN) in the Neck. CBAM dynamically increases the weights of salient channels and spatial regions, thereby enhancing feature representation. The inserted CBAM modules do not alter the input dimensions of subsequent layers or interfere with gradient propagation through residual branches. The channel reduction ratio of CBAM is set to 8 to minimize parameter overhead.

Focal-EIoU Loss is employed in the detection head to compute localization loss at three scales. By adaptively reweighting samples to emphasize high-quality instances, it accelerates the convergence of localization loss during training. Subsequent experiments confirm that this design improves multi-object detection accuracy.

The adopted dual-stage feature extraction mechanism is analogous to the human visual system's processing of complex images: the first stage performs global feature extraction to delineate primary target-recognition regions, while the second focuses on challenging regions to further refine small-target recognition. Rather than enhancing feature extraction at every C3 output module throughout the Backbone, the proposed method applies CBAM only once at the final Backbone output. In the Neck, attention-based enhancement for small-target detection is applied exclusively

after the large-receptive-field (80 × 80) feature output. This design effectively accommodates complex production environments, reduces parameter count and computational complexity, and facilitates embedded deployment.

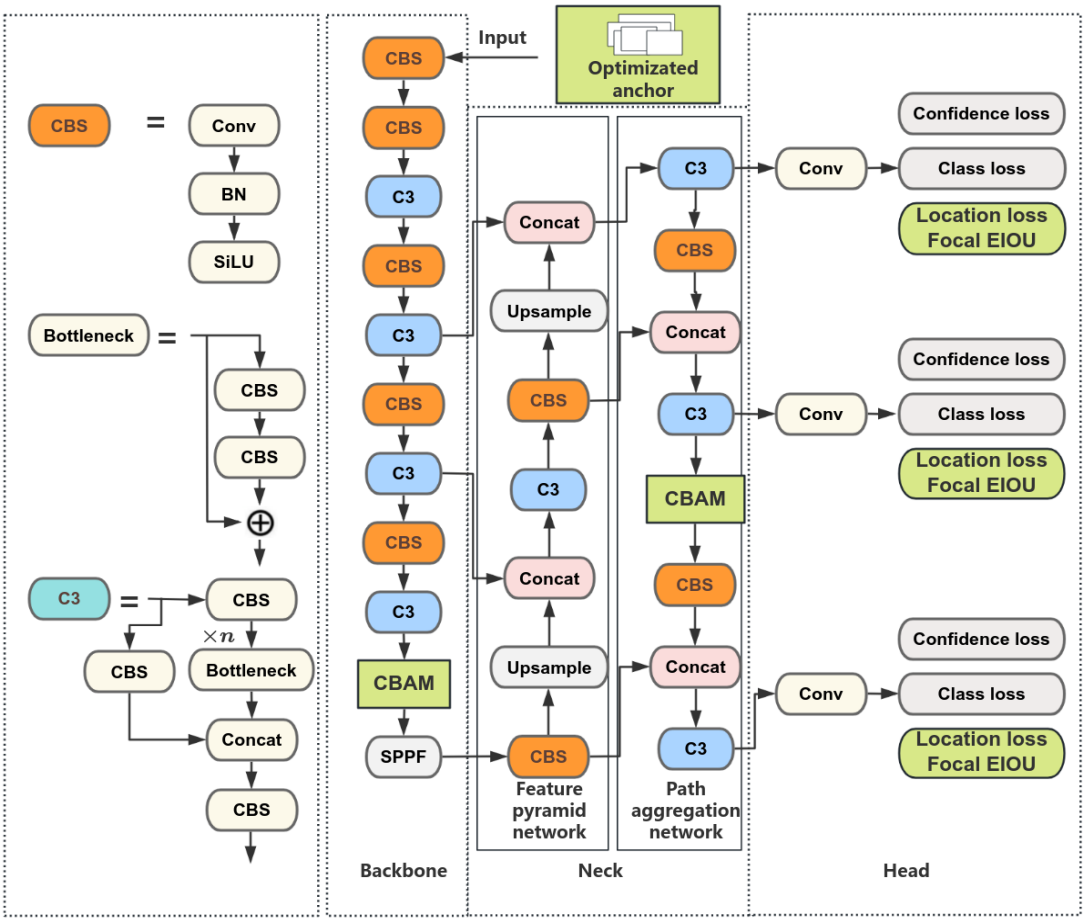

Fig. 2 Improved YOLOv5 network architecture

3.2 Introducing the CBAM attention module

The introduced CBAM is a dual-branch attention mechanism that combines channel and spatial attention. It generates attention information for target features along these two dimensions. The overall network architecture is illustrated in Fig. 3, which shows the integration of CBAM with the residual blocks (ResBlocks) of ResNet. The figure also specifies the exact insertion positions of CBAM within the residual structure of YOLOv5, where CBAM is applied to the convolutional output of each block.

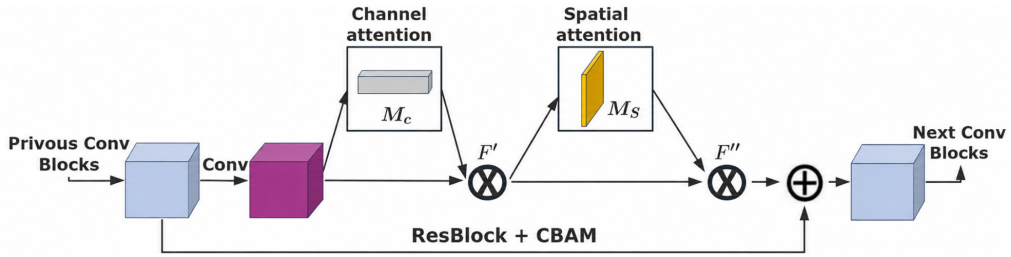

Fig. 3 Integration of CBAM dual-pathway attention structure with a residual block

Given an intermediate feature map $F \in R^{C\times H\times W}$ as input, a 1D channel attention map $M_c \in R^{C\times 1\times 1}$ and a 2D spatial attention map $M_s \in R^{1\times H\times W}$ are sequentially inferred. The entire attention process can be summarized in Eq. 1:

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s(F') \otimes F' \end{aligned} \tag{1}$$

where ⊗ denotes element-wise multiplication. F'' is the final refined output. The convolutional feature attention module of CBAM is shown in Fig. 4.

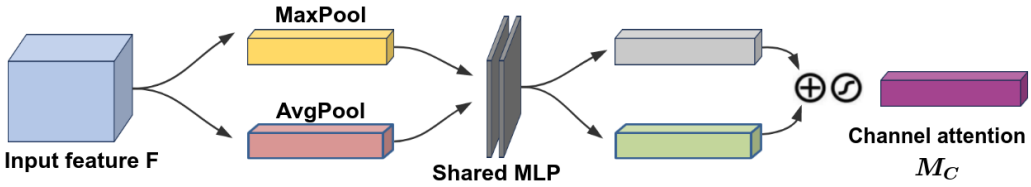

Fig. 4 CBAM channel attention module

For an input feature map F, the module first applies global average pooling across the spatial dimensions to generate a channel-wise vector $V_{\text{avg, C}}$, representing the average activation value for each channel. It then applies global max pooling across the spatial dimensions to generate a vector $V_{\text{max,C}}$, representing the maximum activation value for each channel. These operations aggregate spatial information into channel-only descriptors. $V_{\text{avg, C}}$ and $V_{\text{max, C}}$ are then combined and fed into a shared Multi-Layer Perceptron (MLP) to produce a channel-wise weight vector. This yields the channel attention map M_{C}, whose expression is given in Eq. 2.

$$\begin{aligned} M_{\text{C}} &= \sigma\left(\text{MLP}(V_{\text{max, C}}) + \text{MLP}(V_{\text{avg, C}})\right) \\ &= \sigma\left(W_1\left(W_0(V_{\text{avg, C}})\right) + W_0\left(W_1(V_{\text{max, C}})\right)\right) \end{aligned} \tag{2}$$

where σ denotes the Sigmoid activation function. $V_{\text{max, C}}$ and $V_{\text{avg, C}}$ are $V_{\text{max,C}} = \text{MaxPool}(F)$ and $V_{\text{avg, C}} = \text{AvgPool}(F)$, respectively. $W_0 \in R^{C/r+C}$ and $W_1 \in R^{C+C/r}$ are the weights of the MLP layers, with r denoting the reduction ratio. The spatial attention module of CBAM is shown in Fig. 5.

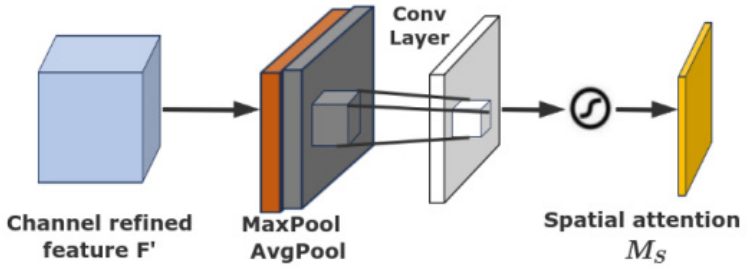

Fig. 5 CBAM spatial attention mechanism

Spatial attention aims to highlight salient objects and suppress background regions. First, max pooling and average pooling are applied to the channels of the input feature map F, retaining only spatial information to obtain $F_{\text{avg, S}}$ and $F_{\text{max, S}}$. These are then concatenated and convolved using a 7 × 7 convolution kernel. The elements of this kernel are initialized to 1, and the kernel does not require iterative learning. The convolved result is then passed through a Sigmoid activation function to generate a spatial weight map, which is subsequently multiplied by the input feature map to obtain the spatial attention map. In summary, spatial attention is computed as shown in Eq. 3.

$$\begin{aligned} M_{\text{S}} &= \sigma\{\text{Conv}_{7\times7}\{[\text{AvgPool}(F); \text{MaxPool}(F)]\}\} \\ &= \sigma(\text{Conv}_{7\times7}\{[F_{\text{avg, S}}; F_{\text{max, S}}]\}) \end{aligned} \tag{3}$$

In automotive battery production, small tools such as scissors exhibit distinct channel and spatial features, making them suitable for feature extraction using attention modules. Raw material recognition also exhibits distinctive spatial characteristics. In the presence of background interference, the spatial attention mechanism enables the model to focus rapidly on material regions while suppressing task-irrelevant features.

3.3 Improved loss function

YOLO-series models typically use CIOU as the bounding-box regression loss function, which combines Intersection over Union (IOU), center-point distance, and aspect-ratio similarity. The formula is given in Eq. 4.

$$L_{\mathrm{CIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{\mathrm{gt}})}{c^2} + \alpha v \tag{4}$$

where ρ represents the Euclidean distance between the center points of the predicted and ground-truth boxes. c is the diagonal length of the smallest enclosing box covering both boxes. v measures aspect-ratio consistency, and α is a weight coefficient. When the aspect ratios are the same but the actual width and height differ substantially, the aspect-ratio term v of CIOU may introduce optimization conflicts. EIOU decouples the aspect-ratio penalty term of CIOU into independent width and height discrepancy terms, enabling direct optimization of the dimensional differences between predicted and ground-truth bounding boxes. The EIOU loss function consists of IOU loss (L_{IoU}), distance loss (L_{dis}), and aspect-ratio loss (L_{asp}). Given a predicted box B and a ground-truth box B^{gt}, where b and b^{gt} represent their respective center points, the loss function is shown in Eq. 5.

$$\begin{aligned} L_{\mathrm{EIoU}} &= L_{\mathrm{IoU}} + L_{\mathrm{dis}} + L_{\mathrm{asp}} \\ &= 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{\mathrm{gt}})}{(w^{\mathrm{c}})^2 + (h^{\mathrm{c}})^2} + \frac{\rho^2(w, w^{\mathrm{gt}})}{(w^{\mathrm{c}})^2} + \frac{\rho^2(h, h^{\mathrm{gt}})}{(h^{\mathrm{c}})^2} \end{aligned} \tag{5}$$

where w^{c} and h^{c} represent the width and height of the smallest enclosing rectangle covering the two boxes. By directly minimizing the width and height differences between the target and anchor boxes, the EIOU loss accelerates convergence and improves localization performance. Its calculation principle is shown in Fig. 6.

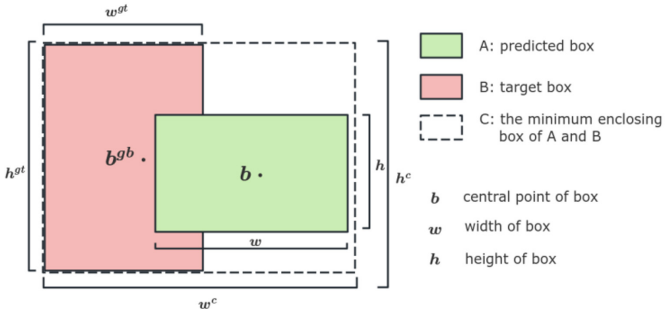

Fig. 6 EIOU position loss calculation principle

Focal Loss was originally developed for classification tasks. It down-weights the loss contribution of easy-to-classify samples, allowing the model to focus on hard samples. Its formula is given in Eq. 6.

$$L_{\mathrm{Focal}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{6}$$

Here, p_t is the predicted probability, and $\gamma \geq 0$ is a modulating factor. When a sample is correctly classified and p_t approaches 1, the weight $(1 - p_t)^{\gamma}$ approaches 0, reducing the loss contribution of simple samples. In the complex battery-production environment, greater weight must be assigned to high-quality samples to improve model learning. This paper introduces Focal-EIOU Loss to further improve the EIOU loss function. During training, this loss places greater emphasis on anchors with higher localization accuracy, thereby improving overall detection precision. The final expression for Focal-EIOU Loss is shown in Eq. 7.

$$L_{\text{Focal-EIOU}} = -\alpha (1 - p)^{\gamma} L_{\mathrm{EIOU}} \tag{7}$$

where α is a class-balance factor that adjusts the weights of positive and negative samples to address class imbalance in the industrial scenario (e.g., if there are few safety-helmet samples, α can be set to 1.8). γ is a hard-example mining factor that increases the weight of difficult examples (e.g., small targets such as scissors) and suppresses the weight of easy examples (e.g., raw-material

types). p is the predicted probability; through $(1-p)^{\gamma}$, it reduces the weight of easy-to-classify samples and focuses learning on difficult examples. L_{EIoU} is the EIoU loss (see Eq. 5), ensuring that the predicted box is closer to the ground-truth box in position, size, and shape.

3.4 Adoption of industrial prior anchor boxes

For YOLOv5's anchor-based mechanism, the prior boxes were originally clustered on the COCO dataset. In this experiment, the anchor aspect ratios were adjusted using K-means clustering with Euclidean distance as the metric. After 100 iterations, five prior anchor-box dimensions were obtained: (16, 20), (65, 75), (102, 206), (202, 211), and (289, 311). These rounded dimensions are relative to the input image scale rather than the grid scale. The small anchors are assigned to small-scale objects, the two intermediate anchors to medium-scale objects, and the large anchors to large-scale objects. The aspect ratios and coordinate distributions of the anchor boxes are shown in Fig. 7.

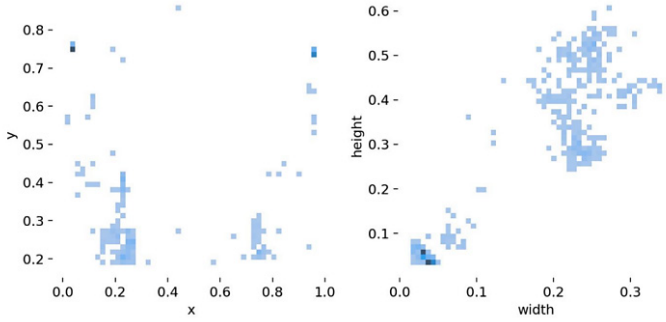

Fig. 7 Anchor aspect ratios and coordinate distributions

4. Experiment

4.1 Experimental environment

The hardware configuration for this experiment included an Intel(R) Core (TM) i7-6700 CPU and an NVIDIA GeForce RTX3060Ti dedicated GPU. The operating system was Windows 10, the programming language was Python 3.9, and the neural network training framework was PyTorch 1.7.0. Training was performed for 100 iterations using input images of 640 × 640 pixels. Network parameters were updated using Stochastic Gradient Descent (SGD), and the hyperparameters are listed in Table 1.

A warm-up training strategy was applied at the beginning of training to mitigate overfitting to mini-batches in the initial phase and prevent model oscillation. The learning rate was subsequently updated using Stochastic Gradient Descent (SGD).

Table 1 Experimental parameters

Parameter	Value
Epoch	100
Batch size	8
Minimum learning rate	0.00001
Maximum learning rate	0.01
Resolution /(pixel × pixel)	640 × 640
Optimizer	SGD

4.2 Dataset

To validate the effectiveness and performance improvements of the proposed method, video data captured by cameras in a factory mixing workshop were selected as sample data. The dataset included monitoring scenes with different backgrounds and scales, as well as variations in lighting conditions, camera angles, weather conditions, and time points. It covered the targets that must be identified during the mixing process, including raw material types, compliance with operator safety requirements, and tool misuse. A total of 2,120 original images were collected and divided into training and testing sets at a 7:3 ratio.

Dataset annotation was performed using the Labelimg tool. The annotated categories primarily included white and black raw materials (with clear markings on the white material), operators' safety helmets, and the presence or absence of scissors on equipment. The annotation of raw material types is shown in Fig. 8, while the number of training instances and prior anchor-box sizes for each target class are shown in Fig. 9.

(a) White raw material (b) Black raw material

Fig. 8 Examples of white material and black material annotations

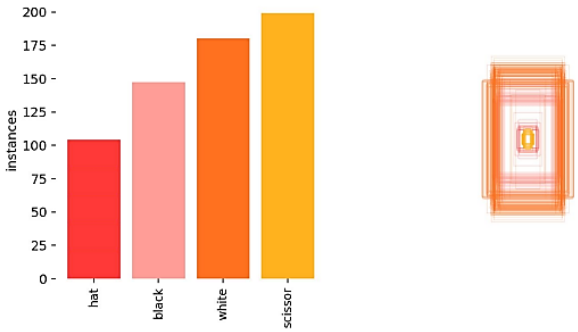

Fig. 9 Number of annotated instances and anchor sizes for training

4.3 Evaluation metrics

This paper employs mAP@0.5 and mAP@0.5:0.95 to evaluate the performance of the proposed network model for multi-object detection in a production environment. *P* (Precision) measures detection accuracy for personnel safety helmets, scissors, and raw materials by representing the proportion of correctly predicted positive samples among all predicted positive samples, as shown in Eq. 8. *R* (Recall) measures the proportion of targets from different classes in the test dataset that are correctly detected, as shown in Eq. 9.

$$P = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP}} \tag{8}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

TP (True Positive) represents the number of samples correctly predicted as positive. FP (False Positive) represents the number of samples incorrectly predicted as positive. For example, an FP occurs when the model predicts that scissors are present in an image although they are not. FN (False Negative) represents the number of samples incorrectly predicted as negative. For example, an FN occurs when the model predicts that no safety helmet is present when one is actually present.

$$\mathrm{AP} = \int_{P}^{1} (R)\, \mathrm{d}R \tag{10}$$

$$\mathrm{mAP} = \frac{1}{N} \sum \mathrm{AP} \tag{11}$$

AP (Average Precision) is the area under the *P*-*R* curve, as shown in Eq. 10. mAP is obtained by averaging the AP values across all classes, where *N* represents the total number of detected classes, as shown in Eq. 11. A higher mAP indicates better overall detection performance and recognition accuracy. AP@0.5 and AP@0.5:0.95 represent AP at an IOU threshold of 0.5 and AP averaged across IOU thresholds from 0.5 to 0.95 in increments of 0.05, respectively. AP reflects detection precision for a specific category, whereas mAP reflects overall detection accuracy across all categories. The mAP@0.5 and mAP@0.5:0.95 metrics are therefore used to evaluate the accuracy and robustness of multi-object detection in battery-production scenarios. In addition, the model's real-time performance is assessed using FPS.

4.4 Experimental results and analysis

After 100 iterations, the model effectively learned the features annotated in the training set, as shown in Fig. 10.

The three loss values, namely coordinate loss, classification loss, and confidence loss, converge rapidly. All three losses approach convergence around the 100th iteration. Notably, the classification loss on the test set approaches 0, supporting the reliability of the model. The overall trends of precision (*P*) and recall (*R*) also converge toward values close to 1 around the 100th iteration.

Some fluctuations in *P* and *R* were observed during certain iterations, indicating background interference in specific scenarios. However, after 100 training iterations, the *P* and *R* metrics converged. During testing, the main precision metrics, mAP@0.5 and mAP@0.5:0.95, also showed good convergence. mAP@0.5 rapidly approached 100 %, while mAP@0.5:0.95 approached 80 %. Fig. 11 illustrates the overall multi-object detection performance, including the detection of small targets such as scissors, the placement of tools on equipment, the use of safety helmets by workers, and the identification of raw material types.

The overall model performance met the expected design objectives. The model detects raw material types with high precision. In cases where safety helmets are not worn, it identifies negative samples and returns confidence scores below 50 %. This supports reliable recognition for subsequent automated analysis and linkage. For partially occluded small targets, the model accurately localizes scissors with a high confidence score of 80 %, thereby supporting compliance with process specifications.

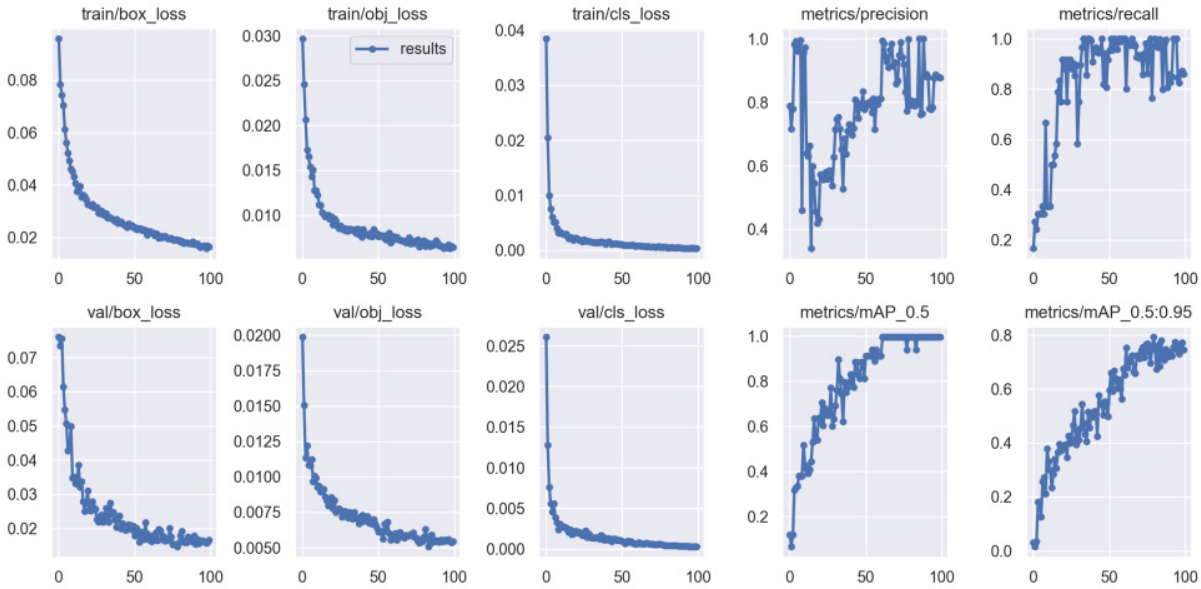

Fig. 10 Loss and accuracy of the improved YOLOv5

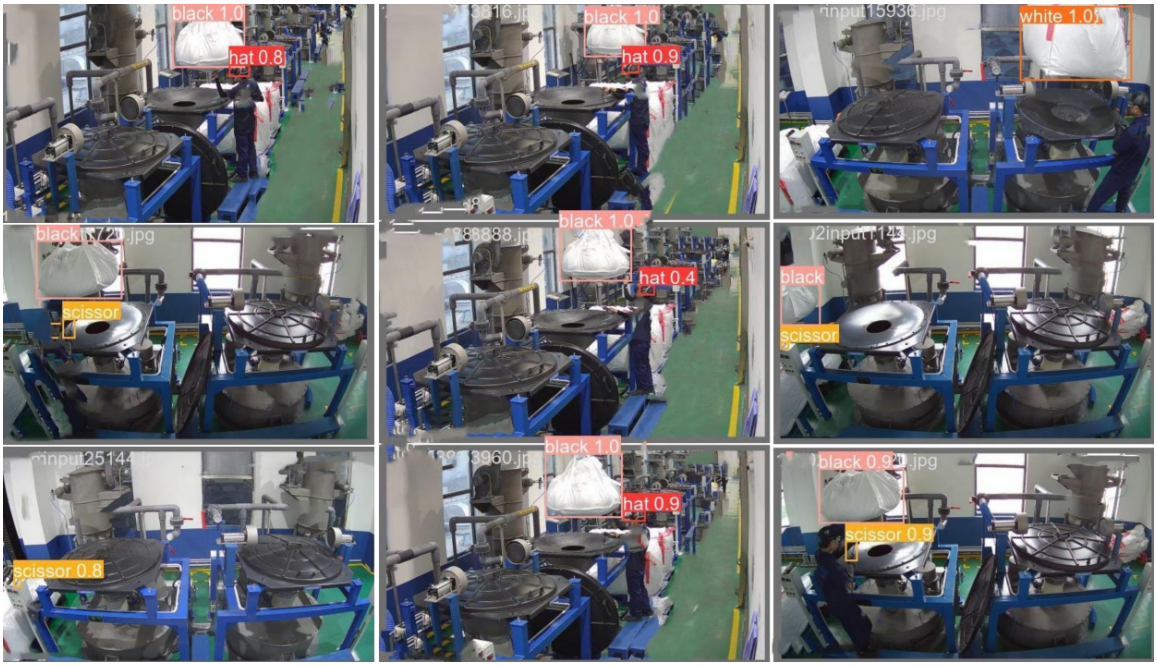

Fig. 11 Multi-object detection in a complex mixing scenario

4.5 Ablation study

To verify the effectiveness of the proposed method, ablation studies were conducted under the same experimental conditions to evaluate the impact of different modules on object detection performance. YOLOv5s was selected as the baseline model, and AP, mAP, *P*, and *R* were used as evaluation metrics.

As shown in Table 2, compared with the original YOLOv5s algorithm, the model introduces the CBAM channel-spatial attention mechanism before the SPPF backbone network to improve small-object detection. The AP@0.5 for scissors increased by 0.6 %, while AP@0.5:0.95 increased by 1.4 %.

Furthermore, introducing the improved weighted loss function into the original model enhanced the model's global receptive field and target classification capability. Consequently, AP@0.5:0.95 increased by 2 %, 2.2 %, and 2.3 % for white material, black material, and safety helmets, respectively.

Finally, integrating both the attention mechanism and the improved loss function into the original model enhanced small-object detection and classification accuracy. The final mAP@0.5 reached 95.2 %, an increase of 2.4 % compared with the original YOLOv5s. The mAP@0.5:0.95 improved by 3.6 %, reaching 78.5 %. Specifically, AP@0.5:0.95 reached 79.4 % for scissors and 75.7 % for safety helmets.

Detection accuracy for safety helmets was relatively lower, which may be related to the number of helmet training samples. Overall, the model achieved improved comprehensive performance when considering both inference speed and detection accuracy, meeting the expected objectives.

Table 2 Ablation experiments

Methods	AP@0.5 (%)				AP@0.5:0.95 (%)							
	White material	Black material	Hat	Scissors	White material	Black material	Hat	Scissors	*P* (%)	*R* (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv5s	96.5	96.4	85.2	93.2	76.4	76.6	71.2	75.3	99.4	99.7	92.8	74.9
YOLOv5s+CBAM	96.5	97.4	85.2	93.8	77.1	77.5	72.4	76.7	99.7	99.6	93.2	75.9
YOLOv5s+Focal-EIoU	97.3	98.4	86.2	94.2	78.4	78.8	73.5	78.1	99.5	99.8	94.0	77.2
Proposed YOLOv5s	97.5	99.4	88.2	95.6	79.2	79.5	75.7	79.4	99.6	99.6	95.2	78.5

4.6 Model comparison experiment

Table 3 compares the improved YOLO model with other lightweight models in the YOLO series. Compared with YOLOv5s, the proposed model's mAP@0.5 increased by 2.6 percentage points. Compared with YOLOv5n, it improved by 3 percentage points. Relative to YOLOv8s and YOLOv8n,

mAP@0.5 increased by 0.4 and 0.7 percentage points, respectively. Overall, these results demonstrate improved accuracy across all tested YOLO variants.

A trade-off was observed: introducing the CBAM attention mechanism reduced FPS. Specifically, the processing speed was 6.7 frames/s lower than that of YOLOv5s. However, this reduction in processing speed is acceptable for practical applications. The proposed model achieves a substantial increase in mAP@0.5 with only a limited increase in time and space complexity. Overall, the proposed model demonstrates the best performance among the compared models.

Table 3 Results of model comparison experiments

Methods	P (%)	R (%)	mAP@0.5 (%)	FPS
YOLOv8n	99.2	97.2	94.4	104.7
YOLOv8s	99.6	98.7	94.7	102.6
YOLOv5n	97.7	94.6	92.1	108.5
YOLOv5s	99.5	94.4	92.5	107.7
Proposed YOLOv5s	99.6	99.2	95.1	101.0

To comprehensively evaluate model performance, visual comparative experiments were conducted for multiple object categories using a single-object recognition paradigm, enabling an intuitive assessment of model capability across different detection scenarios. Fig. 12 presents the detection results of the proposed model alongside YOLOv5 and YOLOv8 in three scenarios: raw material detection, small-object scissor recognition, and safety helmet detection.

In the first scenario, the proposed model successfully identifies raw materials despite background interference, with no prediction failures, achieving a confidence score of 0.96 for white material compared with 0.9 for YOLOv8. In the second scenario, the proposed model achieves a confidence score of 0.9 for small-object scissor recognition, outperforming YOLOv5s at 0.84. In the third scenario, it achieves a confidence score of 0.91 for safety helmet detection, compared with 0.77 for YOLOv5s. These results demonstrate that the proposed model improves detection accuracy over both YOLOv5 and YOLOv8 across multiple object categories.

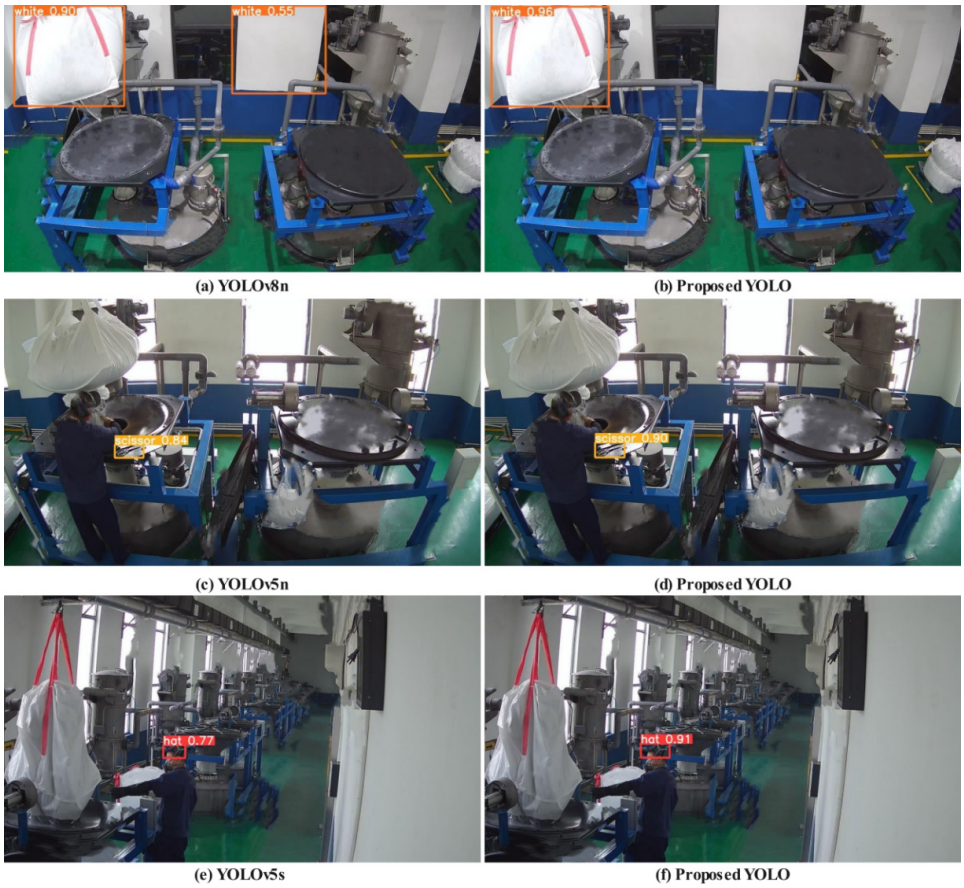

Fig. 12 Visual comparison of detection results between the proposed model and YOLO series versions

The visualizations further demonstrate the superior robustness of the proposed model. In cases (a) and (b), which involve complex background interference, YOLOv8 yields a low confidence score of 0.55 and shows noticeable sensitivity to background clutter, whereas the proposed model effectively suppresses interference, produces no false detections, and achieves higher confidence. Cases (c) and (d) involve specular reflections under varying illumination, and the proposed model achieves higher recognition accuracy for small objects under strong highlights. Cases (e) and (f) contain motion-blurred targets observed from different camera viewpoints, for which the proposed model achieves superior recognition accuracy for helmets worn by moving personnel.

4.7 Performance and deployment

With an input resolution of 640 × 640 pixels in the battery-production scenario, the improved YOLOv5 model contains 7.48 MB trained parameters, has an activated memory footprint of 16.89 MB and a peak memory consumption of 27.65 MB, and requires 11.23 GFLOPs. These values are considerably lower than those of other YOLOv5 variants and the YOLOv8 series, making the model well suited to embedded deployment.

5. Conclusion

Industrial production environments require visual detection models that combine high precision, low latency, and lightweight architecture. When deployed on embedded devices at production sites, the recognition signals generated by these models can be integrated with production automation signals to support on-site management strategies. Accordingly, integrating computer vision detection models with other production systems is an important research direction for intelligent manufacturing. To address the key challenges of dynamic detection in complex scenes, multi-scale and small-object recognition, and environmental interference in automotive battery manufacturing, this paper proposes a lightweight, improved YOLOv5 detection model. By incorporating attention mechanisms and an improved loss function, the proposed model significantly enhances detection accuracy in complex environments. Comparative experiments further demonstrate its robustness under environmental interference, occlusion, strong specular reflections, and motion blur.

- Inspired by the two-stage processing of visual features in the human brain, the proposed approach combines global feature scanning with focused processing of difficult cases. Through lightweight optimization, a two-stage feature enhancement network is constructed. By embedding CBAM attention modules in the Backbone and Neck of YOLOv5, the model dynamically focuses on salient regional features while suppressing background noise. Experimental results show that mAP@0.5 reaches 93.2 %, indicating improved recognition accuracy. Because only two attention modules are employed, memory consumption and computational complexity are lower than in approaches using multiple attention stages.
- Industry-specific anchor parameters generated through K-means clustering address the mismatch between YOLOv5's default anchor sizes and the geometric characteristics of industrial objects. Replacing conventional CIoU with Focal-EIoU Loss introduces adjustable penalty terms for sample weighting and aspect-ratio discrepancy. Ablation experiments show that mAP@0.5 reaches 94.0 %, supporting the effectiveness of the combined dynamic loss function and anchor optimization.

Although the proposed model demonstrates promising performance in the battery-production environment, its feature extraction capability under complex optical conditions remains limited. Future work should introduce a contrast-enhancement mechanism at the preprocessing stage to improve input image quality and strengthen model robustness. Additional data from diverse industrial scenarios should also be incorporated to further validate and improve the model's generalization capability. Finally, embedded deployment should be completed to enable dynamic compliance monitoring of personnel, equipment, and the environment using 3D video streams captured by on-site cameras.

The improved model can also be extended to applications such as automotive component manufacturing and full-vehicle assembly, contributing to the continued advancement of quality control in smart factories.

References

[1] Xu, J., Sun, Q., Han, Q.-L., Tang, Y. (2025). When embodied AI meets Industry 5.0: Human-centered smart manufacturing, *IEEE/CAA Journal of Automatica Sinica,* Vol. 12, No. 3, 485-501, doi: 10.1109/JAS.2025.125327.

[2] Zhou, M.C. (2022). Editorial: Evolution from AI, IoT and big data analytics to metaverse, *IEEE/CAA Journal of Automatica Sinica,* Vol. 9, No. 12, 2041-2042, doi: 10.1109/JAS.2022.106100.

[3] Duan, J., Yu, S., Tan, H.L., Zhu, H., Tan, C. (2022). A survey of embodied AI: From simulators to research tasks, *IEEE Transactions on Emerging Topics in Computational Intelligence,* Vol. 6, No. 2, 230-244, doi: 10.1109/TETCI.2022.3141105.

[4] Song, Y., Sun, P., Liu, H., Li, Z., Song, W., Xiao, Y., Zhou, X. (2024). Scene-driven multimodal knowledge graph construction for embodied AI, *IEEE Transactions on Knowledge and Data Engineering,* Vol. 36, No. 11, 6962-6976, doi: 10.1109/TKDE.2024.3399746.

[5] Ren, S., He, K., Girshick, R., Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 39, No. 6, 1137-1149, doi: 10.1109/tpami.2016.2577031.

[6] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C. (2016). SSD: Single shot multibox detector, In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.), *Computer vision – ECCV 2016. ECCV 2016. Lecture notes in computer science,* Vol. 9905, Springer, Cham, Switzerland, 21-37, doi: 10.1007/978-3-319-46448-0_2.

[7] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection, In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, Nevada, USA, 779-788, doi: 10.1109/CVPR.2016.91.

[8] Yin, Q., Yang, W., Ran, M., Wang, S. (2021). FD-SSD: An improved SSD object detection algorithm based on feature fusion and dilated convolution, *Signal Processing: Image Communication*, Vol. 98, Article No. 116402, doi: 10.1016/j.image.2021.116402.

[9] Moon, J., Son, M., Oh, B., Jin, J., Shin, Y. (2024). Automatic voltage stabilization system for substation using deep learning, *Computer Science and Information Systems*, Vol. 21, No. 2, 437-452, doi: 10.2298/CSIS220509050M.

[10] Pang, J.L. (2023). Adaptive fault prediction and maintenance in production lines using deep learning, *International Journal of Simulation Modelling,* Vol. 22, No. 4, 734-745, doi: 10.2507/IJSIMM22-4-CO20.

[11] Jian, Z.-J., Su, M.-H. (2024). Surface defect detection in industrial parts using the YOLO-based method with enhanced spatial features and attention mechanism, In: *Proceedings of 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, Takamatsu, Japan, 511-515, doi: 10.1109/IIAI-AAI63651.2024.00098.

[12] Hsu, W.-Y., Lin, W.-Y. (2021). Ratio-and-scale-aware YOLO for pedestrian detection, *IEEE Transactions on Image Processing*, Vol. 30, 934-947, doi: 10.1109/tip.2020.3039574.

[13] Badmos, O., Kopp, A., Bernthaler, T., Schneider, G. (2020). Image-based defect detection in lithium-ion battery electrode using convolutional neural networks, *Journal of Intelligent Manufacturing*, Vol. 31, No. 4, 885-897, doi: 10.1007/s10845-019-01484-x.

[14] Lee, J., Hwang, K. (2022). YOLO with adaptive frame control for real-time object detection applications, *Multimedia Tools and Applications*, Vol. 81, No. 25, 36375-36396, doi: 10.1007/s11042-021-11480-0.

[15] Marco-Detchart, C., Rincon, J.A., Carrascosa, C., Julian, V. (2024). Evaluation of deep learning techniques for plant disease detection, *Computer Science and Information Systems*, Vol. 21, No. 1, 223-243, doi: 10.2298/CSIS221222073M.

[16] Zhao, Q., Fang, S., Li, Y., Shang, H., Zhang, H., Shen, T. (2025). Track defect detection based on improved YOLOv5s, *IEEE Transactions on Industrial Informatics*, Vol. 21, No. 4, 3346-3355, doi: 10.1109/TII.2024.3523593.

[17] Liu, J., Zhou, G. (2024). Learning discriminative representations through an attention mechanism for image-based person re-identification, *Computer Science and Information Systems*, Vol. 21, No. 4, 1483-1498, doi: 10.2298/CSIS230829044L.

[18] Zhu, G., Liu, Y., Wang, J. (2024). Multi-frame network feature fusion model and self-attention mechanism for vehicle lane line detection, *Computer Science and Information Systems*, Vol. 21, No. 4, 1699-1723, doi: 10.2298/CSIS240314054Z.

[19] Lv, Q.H., Chen, J., Chen, P., Xun, Q.F., Gao, L. (2024). Flexible Job-shop Scheduling Problem with parallel operations using Reinforcement Learning: An approach based on Heterogeneous Graph Attention Networks, *Advances in Production Engineering & Management*, Vol. 19, No. 2, 157-181, doi: 10.14743/apem2024.2.499.

[20] Yang, J., Zhang, Z., Li, W., Wang, X., Yang, S., Yang, C. (2025). Underwater acoustic target classification using auditory fusion features and efficient convolutional attention network, *IEEE Sensors Letters*, Vol. 9, No. 3, Article No. 7001304, 1-4, doi: 10.1109/LSENS.2025.3541593.

[21] Zhou, K., Haimudula, A., Tang, W. (2024). Dual-branch convolution network with efficient channel attention for EEG-based motor imagery classification, *IEEE Access*, Vol. 12, 74930-74943, doi: 10.1109/ACCESS.2024.3404634.

[22] Wei, S., Shen, S., Liu, D., Song, Y., Gao, R., Wang, C. (2024). Coordinate attention enhanced adaptive spatiotemporal convolutional networks for traffic flow forecasting, *IEEE Access*, Vol. 12, 140611-140627, doi: 10.1109/ACCESS.2024.3460405.

[23] Zhang, Y.-P., Zhang, Q., Kang, L., Luo, Y., Zhang, L. (2022). End-to-end recognition of similar space cone–cylinder targets based on complex-valued coordinate attention networks, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, 1-14, Article No. 5106214, doi: 10.1109/TGRS.2021.3115624.
[24] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S. (2018). CBAM: Convolutional block attention module, In: *Proceedings of the European Conference on Computer Vision (ECCV)*, 3-19, doi: 10.1007/978-3-030-01234-2_1.
[25] Zong, H., Shuai, B. (2025). Manufacturing process quality prediction via temporal knowledge graph reasoning with adaptive multi-scale temporal path fusion and self-attention mechanism, *Advances in Production Engineering & Management*, Vol. 20, No. 3, 380-390, doi: 10.14743/apem2025.3.547.
[26] Ding, L., Wang, L., Wang, Y., Yu, S., Xiao, J. (2024). A text recognition algorithm based on a dual-attention mechanism in complex driving environment, *Tehnički Vjesnik – Technical Gazette,* Vol. 31, No. 1, 247-253, doi: 10.17559/TV-20231023001052.
[27] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale, *arXiv*, doi: 10.48550/arXiv.2010.11929.
[28] Hao, H., Yu, X. (2024). YOLOv8-UCB: Visual detection of pouch battery using improved YOLOv8, *IEEE Access*, Vol. 12, 194899-194910, doi: 10.1109/ACCESS.2024.3520181.
[29] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D. (2020). Distance-IoU loss: Faster and better learning for bounding box regression, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 7, 12993-13000, doi: 10.1609/aaai.v34i07.6999.
[30] Zheng, H., Gao, X., Yang, X., Jing, G., Yang, M., Liu, Y. (2025). A multi-objective feature selection and self-paced ensemble framework for semiconductor defect detection, *Advances in Production Engineering & Management*, Vol. 20, No. 4, 458-474, doi: 10.14743/apem2025.4.552.
[31] Wang, S.-Y., Qu, Z., Gao, L.-Y. (2024). Multi-spatial pyramid feature and optimizing focal loss function for object detection, *IEEE Transactions on Intelligent Vehicles*, Vol. 9, No. 1, 1054-1065, doi: 10.1109/TIV.2023.3282996.