SCIPEDIA Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

Revista Internacional Métodos numéricos para cálculo y diseño en ingeniería RIMNI

In cooperation with CIMNE

# Detection of Small Rod-End Joint Bearings via Deep Feature Fusion and Confidence Propagation Clustering

**Jinmin Peng[1,2], Ruifeng Ye[3,*], Song Lan[4,*], Tenghao Xiao[1,2], Chen Xu[1,2] and Yancong Song[1,2]**

[1]Fujian Key Laboratory of Intelligent Processing Technology and Equipment, Fujian University of Technology, Fuzhou, 350118, China

[2]School of Mechanical and Automotive Engineering, Fujian University of Technology, Fuzhou, 350118, China

[3]College of Artificial Intelligence, Yango University, Fuzhou, 350015, China

[4]College of Automation Engineering, Fujian College of Water Conservancy and Electric Power, Yong'an, 366000, China

**ABSTRACT**

Machine vision is used to detect dense, small rod-end joint bearings in sliding ball surfaces with little feature information and high variability. However, this leads to inaccurate identification, affecting production efficiency. This study proposes a deep-learning object-detection algorithm model that allows the network to retain more semantic information. We introduced the space-to-depth convolution (SPD-Conv) step-free convolution module to improve the backbone network and developed a multi-level feature fused SPD (MFSPD) deep feature fusion module to redesign the neck network to improve the feature extraction ability and detection accuracy for small targets. Furthermore, we added a small P4 detection head in the head network (i.e., prior box acquisition on the dataset using the weighted k-means algorithm), increased the matching degree of the prior box and feature layer, and accelerated the model convergence. To improve the confidence propagation clustering (CP-Cluster) analysis algorithm for post-processing, we optimized the prediction box confidence degree and detection speed. The algorithm performance was evaluated on homemade, T-LESS, and COCO datasets. The mAP@.5 values of the target detection algorithm for the homemade and T-LESS datasets were 96.9% and 93.8%, respectively, and the mAP was 55.9% for the COCO dataset. The experimental results indicate that the algorithm has a high detection accuracy and good feature extraction ability. Thus, it has considerable advantages for small-object detection and provides a reference for the detection of small parts.

OPEN ACCESS

**Accepted:** 27/09/2024

**Submited:** 24/07/2024

**DOI**
110.23967/
j.rimni.2024.10.56511

**Keywords:** Small size; rod-end joint bearing; MFSPD; improved CP-Cluster; weighted k-means

## 1 Introduction

Rod-end joint bearings are indispensable in mechanical equipment, acting analogously to joints in the human body. Owing to their large variety and minute differences [1], sorting bearings via manual detection is inefficient and has low accuracy. Therefore, machine vision and robotics have been introduced in object detection tasks [2]. Machine-vision-guided robotics can significantly increase

detection accuracy and efficiency. However, in CNC machine tool production lines employing a hybrid manufacturing approach with varying models and sizes, traditional vision-based detection of small bearings remains challenging due to the complex industrial environment. Issues such as surface reflections, disordered stacking, and the high variability of the sliding ball surfaces often result in insufficient feature information. To overcome these challenges, we have employed deep learning techniques. The small size of the rod-end bearings is illustrated in Fig. 1.

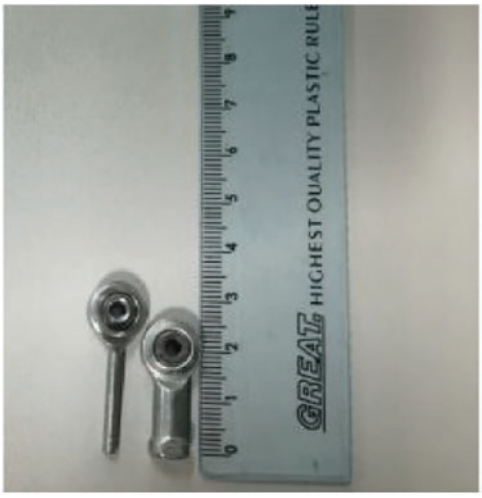


**Figure 1:** Small rod-end joint bearing

Current object detection algorithms based on deep learning can be broadly categorized into two-stage [3–6] and one-stage [7–10] approaches. Kim et al. [11] demonstrated that integrating super-resolution (SR) with object detection significantly improves accuracy for small objects by enhancing image resolution, thereby reducing false detections. Wahyudi et al. [12] emphasized the effectiveness of strategies such as multiscale feature fusion and contextual information enhancement in addressing the limitations of small-object detection. Further more, Park et al. [13] introduced a multimodal data fusion approach that mitigates the challenges of noise and low resolution, further improving detection performance in complex environments. Despite these advancements, several limitations persist in small object detection:

1. Insufficient feature representation: Existing deep learning models struggle with inadequate feature information for small objects due to limited spatial resolution and multiple downsampling operations [14,15]. The large receptive fields and inherent low-resolution characteristics of these networks further exacerbate the challenge of extracting discriminative features for small objects, leading to increased false positives and false negatives.

2. Suboptimal utilization of contextual information: While contextual cues are crucial for small object detection, current methods often fall short in effectively integrating multi-scale feature fusion and contextual information enhancement [16,17]. This deficiency hinders the model's ability to comprehend small objects within their environmental context, particularly in complex scenes or under occlusion.

3. Limitations of Non-Maximum Suppression for dense small objects: Conventional Non-Maximum Suppression (NMS) algorithms face challenges when processing densely clustered small objects. The fixed IoU threshold in standard NMS leads to suboptimal performance in scenarios

with tightly packed small targets [18,19]. This constraint results in decreased recall rates and reduced localization accuracy, especially in complex scenes where small objects are in close proximity or partially occlude each other.

To address these limitations and achieve the required detection accuracy for small rod-end joint bearings, we propose several improvements to the existing object detection frameworks. Our approach focuses on enhancing feature extraction capabilities, optimizing feature fusion, and refining post-processing techniques.

Among the current state-of-the-art object detection algorithms, the YOLO series has shown promising results [20]. While YOLOv6 [21] and YOLOv7 [22] demonstrate superior detection accuracy and model performance compared to YOLOv5, they incur higher computational costs in terms of floating point operations (FLOPs) and parameter counts. In practical applications, particularly for small object detection tasks, the YOLOv5 architecture offers a favorable balance between performance and efficiency. YOLOv5, an improved version of YOLOv4 [23], exhibits significantly higher detection speed and accuracy [24], while providing models with various trade-off options for accuracy and speed.

Building upon these considerations, we based our study on YOLOv5 and proposed the SCP-YOLOv5 object detection algorithm. Our approach incorporates enhanced feature fusion modules and an improved confidence propagation clustering algorithm to address the aforementioned limitations of small object detection. The detailed improvements and methodology of our proposed SCP-YOLOv5 algorithm are presented in Section 2.

To validate the effectiveness of our proposed method, we developed a specialized rod-end joint bearing dataset and evaluated the algorithm's performance using this custom dataset along with the established T-LESS and COCO benchmarks.

## 2 Methods

The production and classification of rod-end bearings involve complex situations, such as stacking and occlusion, resulting in insufficient feature information. SCP-YOLOv5 addresses the aforementioned issues by making multiple improvements to the backbone, neck, and head to enhance feature extraction and strengthen representation capability. In the backbone, more feature map outputs are added, and the SPD-Conv [25] module is introduced to improve the network structure and feature extraction capabilities. In the neck, the MFSPD module based on SPD-Conv is proposed to redesign the structure for better retention of the fused feature information. In the head, the P4 detection head is added, and weighted k-means [26] clustering is utilized on the rod-end bearing data to obtain suitable anchors, increasing the detection accuracy and convergence speed for small rod-end bearings. Finally, the confidence propagation clustering (CP-Cluster) algorithm [27] is further enhanced in the post-processing stage. This improvement effectively addresses the limitations of traditional NMS methods, particularly in mitigating the issue of detection degradation caused by excessive bounding box overlap. Through these improvements, SCP-YOLOv5 solves problems in rod-end bearing detection and enhances the detection performance. This strengthens the detection and classification capabilities for small targets.

### *2.1 Backbone Design*

Existing CNN module architectures use strided convolutions or pooling layers in series or parallel, which cause loss of fine-grained texture information and less efficient feature learning for rod-end joint bearings—particularly small ones. To address these problems, we introduced an SPD-Conv module into the backbone. The module contains a space-to-depth layer and non-strided convolution layer.

The space-to-depth layer uses an unprocessed image transformation technique [28] to split the rod-end joint bearing feature maps from spatial dimensions to concatenated channel dimensions, as shown in Fig. 2.

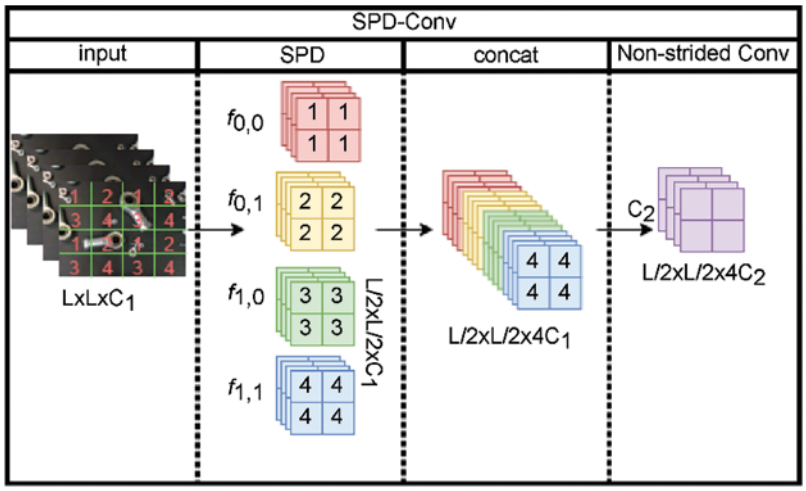


**Figure 2:** SPD-Con v module structure

The backbone is based on the improved CSPDar knet53 with an additional 160 × 160 × 128 feature map output for subsequent small-object detection. All strided convolution layers are replaced with SPD-Con v modules. The improved architecture outputs four feature maps with scales of 160 × 160 × 128, 80 × 80 × 256, 40 × 40 × 512, and 20 × 20 × 1024. They are then input to the neck for further fusion and enhancement, as shown in Fig. 3.

### *2.2 Neck Design*

For accurately detecting small rod-end bearings, a P4 detection head was added. The increased depth of the neck provides support for the changes in the head, while the increased depth raises the risk of long-term memorization deficiency. To retain more feature information and enhance the feature intensity of small rod-end bearings, the module had to be able to retain more information and fuse features to increase the feature value intensity for each dimension to contain more feature information. According to this concept, a deep-feature fusion module called the MFSPD based on the SPD module was developed. Its structure is shown in Fig. 4.

### *2.3 Head Design*

#### *2.3.1 Detection Head Design*

The head of the rod-end bearing has high variability. The same joint bearing can exhibit different sliding spherical surfaces, as shown in Fig. 5. The head network of YOLOv5 contained only P8, P16, and P32 detection heads. To solve the problem of reduced detection accuracy caused by the aforementioned factors and adapt the model to the detection of small rod-end bearings, a P4 detection head was added to the head to output a 160 × 160 feature map.

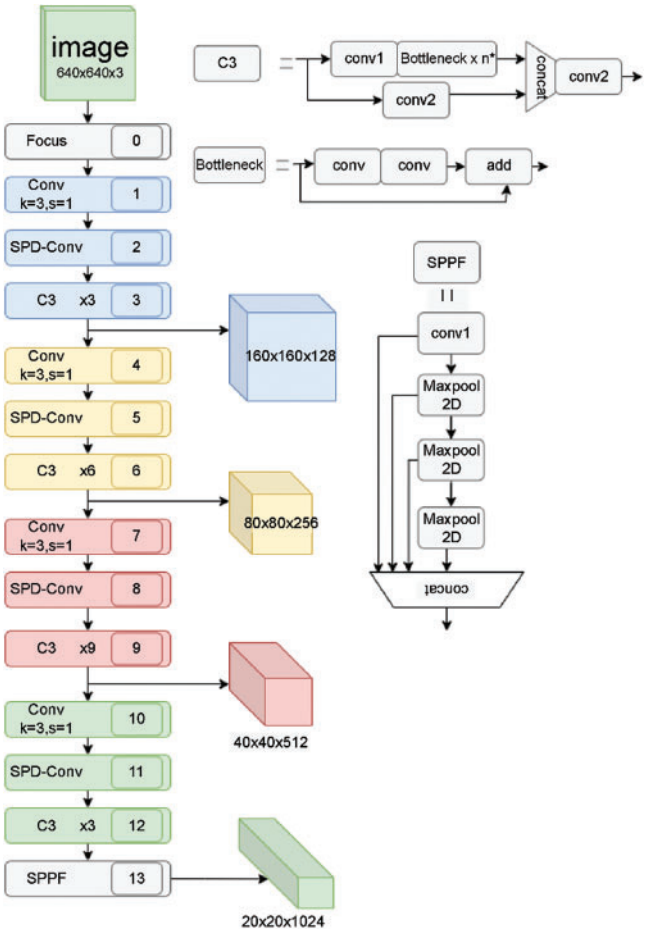


**Figure 3:** Backbone structure: The novel SPD-Conv module and additional feature map output channel address limitations in existing CNN architectures. This design preserves fine-grained texture information and enhances feature learning efficiency for rod-end joint bearings, particularly beneficial for smaller specimens. The structure mitigates information loss typically associated with strided convolutions or pooling layers, enabling more accurate and detailed analysis across various bearing sizes

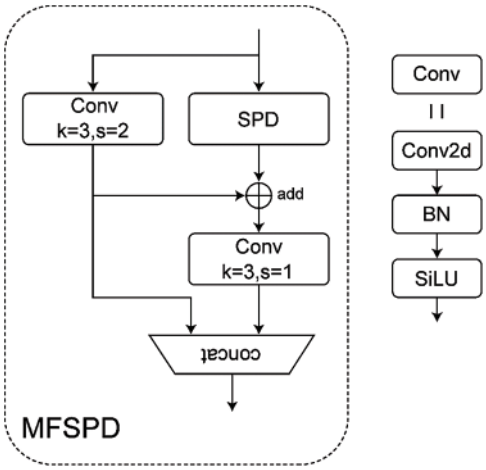


**Figure 4:** MFSPD structure: The MFSPD , based on the SPD module, is designed to retain more detailed feature information and intensify feature values across dimensions. This structure mitigates the risk of long-term memorization deficiency while improving the detection accuracy of small rod-end bearings through advanced feature fusion and retention techniques

**Figure 5:** Different angles of sliding spheres

### *2.3.2 Obtaining Prior Boxes*

The dataset anchor had to be determined before training. Suitable anchors can increase the detection accuracy and accelerate the convergence of the model. The anchor of YOLOv5 was obtained from the COCO dataset, whereas the rod-end bearing dataset had small sizes and dense and mixed cluttered stacking, resulting in an anchor mismatch with the dataset. To solve these problems, re-clustering the annotation boxes and obtaining suitable anchors were necessary. In this study, a weighted k-means algorithm based on the original k-means algorithm is proposed, in which a weight coefficient is introduced for each sample. According to the chaotic detection state of the rod-end bearing, the maximum IoU of the cluster centers is used to evaluate the clustering results. The steps of the algorithm are as follows:

(1) The clustering center K is adaptively adjusted according to the class label file of the dataset.

(2) The distance from each sample to the center point of the feature map is calculated using the following formula:

$$d_{ic} = \sqrt{\left(x_i - \hat{x}_c\right)^2 + \left(y_i - \hat{y}_c\right)^2}, \tag{1}$$

where $x_i$ and $y_i$ denote the $i$ horizontal and ordinate samples, respectively, and $\hat{x}_c$ and $\hat{y}_c$ denote the $c$ central horizontal and ordinate samples, respectively.

(3) The weight of the cluster center of each sample is considered, and the weight matrix is constructed. Eq. (2) gives the mean value of the distances between all samples and a central point $a$, and Eq. (3) is the weight calculation formula.

$$\overline{d}_a = \frac{1}{n}\sum_{i=1}^{n} d_{ic}, \tag{2}$$

$$w_{ijc} = \max\left(0, -\frac{d_{ic} - \overline{d}_c}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(d_{ic} - \overline{d}_c\right)^2}}\right). \tag{3}$$

Here, $w_{ijc}$ is the weight of the $i$ sample and $j$ cluster center, and $c$ is the feature map center point corresponding to the cluster center.

(4) The target function $D$, i.e., the cluster center, is updated:

$$D = \min\sum_{i=1}^{n}\sum_{j=1}^{k}\left[1 - w_{ijc}\frac{B_i \cap C_j}{B_i \cup C_j}\right], \tag{4}$$

where $B_i$ represents the width of the annotation box for sample $i$, and $C_j$ denotes the preselection box for the $j$ cluster center.

(5) All the cluster centers are iteratively updated until the cluster-center position is constant.

### *2.4 Overall Structure of Improved Algorithm*

According to the improvements and designs of the aforementioned structures, the overall structure of the SCP-YOLOv5 algorithm is proposed, as shown in Fig. 6. The input image size is 640 × 640 × 3 pixels, and the backbone outputs four feature maps of different sizes to the neck for further enhancement and fusion of features. Finally, the head performs target-tag classification and prediction-box regression.

A feature pyramid network (FPN) is used in the neck. Upsampling is performed from bottom to top, and convolution pooling is performed from top to bottom to fuse feature maps at different levels. This increases the feature value intensity and improves semantic features to enhance the multiscale target detection capabilities.

In the feature pyramid, deeper feature maps contain higher-level semantic features, whereas shallower feature maps contain more target location information. Layer 25 fuses 4 scale feature maps and outputs a 160 × 160 × 128 feature map. The output is divided into two paths: one output is input to the P4 detection head and the other is used to enhance the target position information from top to bottom.

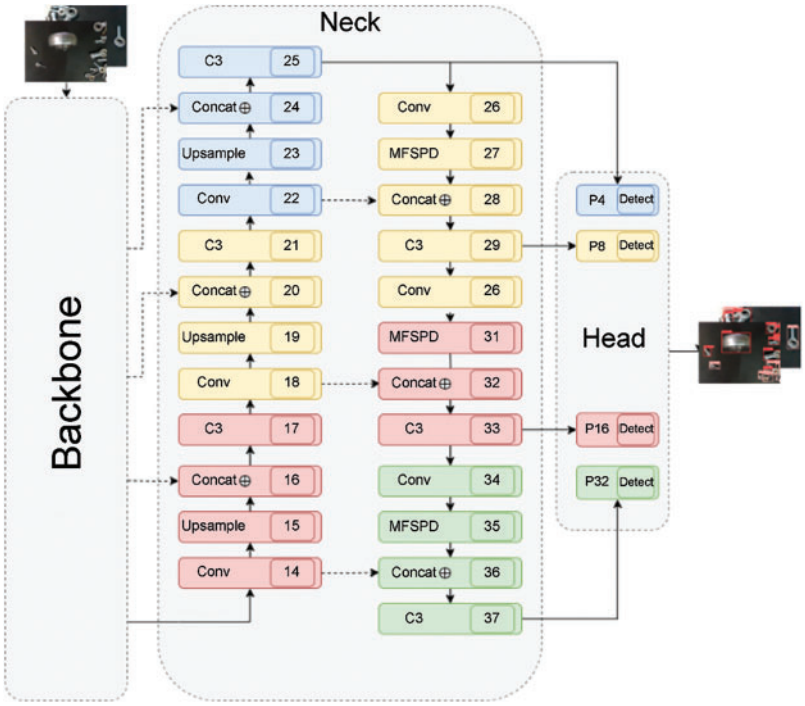


**Figure 6:** SCP-YOLOv5 structure

### *2.5 Processing of Detection Results*

Considering the defects of the traditional Non-Maximum Suppression (NMS) algorithm, the prediction box with the highest confidence is not necessarily the best prediction box; sorting is required before sequential processing. To address such issues, the CP-Cluster algorithm [28] automatically propagates messages between adjacent prediction boxes. This helps adjust the confidence after multiple iterations and improve strong confidence areas while reducing weak ones, which allows fully parallel processing, increasing the algorithm processing speed. Furthermore, rod-end bearings are prone to situations in which the prediction boxes severely overlap when stacked. Therefore, the IoU is not suitable as an evaluation metric. The distance IoU (DIoU) [29] was introduced to replace the IoU as a parameter for evaluating the boundary-box location. When prediction boxes overlap, negative message passing suppression is more reasonable to ensure accurate prediction of occluded objects in clusters. When prediction boxes are non-intersecting, the weakest friends can still provide the most precise movement directions for the strongest friend. The improved scheme combines the two algorithms to resolve the speed and overlap issues of the NMS algorithm. In Fig. 7, the detection results of the CP-Cluster algorithm and the improved scheme are shown on the left and right, respectively.

As shown in Fig. 8, all the candidate boxes are first converted into an undirected graph set, and then the candidate boxes pass positive and negative messages to each other in the graph. Finally, the discarded candidate boxes are eliminated, and the confidence of the selected candidate boxes is increased.

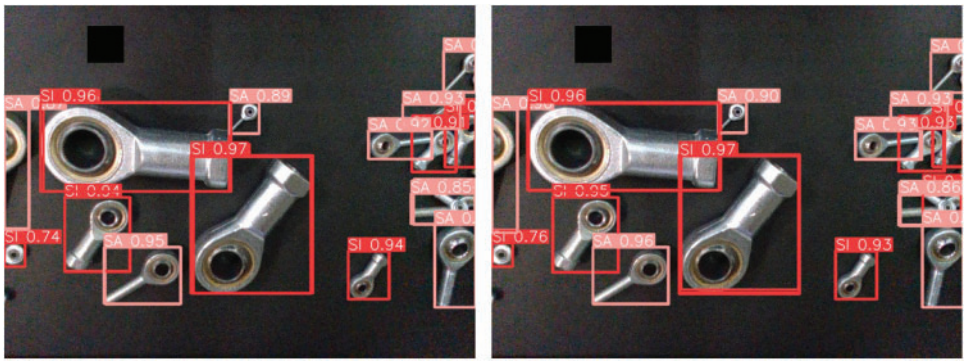

**Figure 7:** Detection results of the CP-Cluster algorithm (left) and the improved scheme (right)

**Figure 8:** CP-Cluster processing flow diagram

In positive message passing, friend clusters with lower confidence and an IoU higher than a specific threshold are judged as weaker. The confidence of the strong side is updated according to the number and confidence of weaker friends, as follows:

$$M_p(i) \leftarrow \frac{Q}{Q+1} * \left(1 - \hat{P}(b_i)\right) * \max_{\hat{b} \in W_{b_i}} \hat{P}\left(\hat{b}\right), \tag{5}$$

where $w_{b_i}$ denotes the set of weak friends, and $P$ denotes the degree of confidence.

In the spread of negative news, the original ranking from high to low changes to a graph structure, eliminating the ranking and suppressing it twice. Additionally, an SUP matrix is added to prevent the prediction box from being repeatedly suppressed by the same prediction box, and the role of hyperparameter $\zeta$ is to limit this phenomenon.

Eq. (6) is suppressed by the strongest friend, and Eq. (7) updates the weak-side confidence.

$$T_{(b_j,b_i)} \leftarrow \alpha * \frac{\hat{P}(b_j)}{\hat{P}(b_i)} + (1-\alpha) * \frac{DIoU(b_j, b_i)}{\theta}, \tag{6}$$

$$M_n(i) \leftarrow \hat{P}(b_i) * DIoU\left(b_i, \underset{b_j \in N_{b_j}, SUP_{j,i} \le \zeta}{\arg\max}\ T_{(b_j,b_i)}\right), \tag{7}$$

$$DIoU(b_j, b_i) = 1 - IoU + \frac{\rho^2(b_j, b_i)}{c^2}. \tag{8}$$

Eq. (8) gives the DIoU evaluation parameters. Here, $\rho$ represents the Euclidean distance between two center points, and $c$ represents the diagonal of the minimum rectangle that can cover both the anchor and the target box.

Through iterative cycles, the algorithm converges to the optimal solution. Finally, prediction boxes with low confidence are filtered out, leaving only the prediction boxes with the highest confidence.

### *2.6 Experimental Environment and Model Training*

#### *2.6.1 Experimental Configuration*

Details regarding the configuration of the experimental environment are presented in Table 1.

**Table 1:** Experimental environment configuration

| Software and hardware environment | Model parameters |
|---|---|
| OS | Ubuntu 20.04 |
| Development language | Python 3.7.0 |
| Deep-learning framework | PyTorch 1.8 |
| CUDA | CUDA version 11.1 |
| IDE | PyCharm2022.1.3 +Anaconda3 |
| GPU | NVIDIA GeForce RTX3050Ti |
| CPU | AMD Ryzen7 5800h 16G |
| Data processing tool | OpenCV 4.5.1.48 |
| Camera | Intel REALSENSE D435I |
| Data annotation tool | LabelImg |

#### *2.6.2 Experimental Dataset*

Using a D435I camera, 1110 rod-end bearing images with a resolution of 640 × 480 were collected by calling the SDK. Data annotation was performed using LabelImg with the labels divided into two categories: SI and SA. The lower left part of Fig. 9a shows the SI inner row rod-end bearing, and the upper right part shows the SA outer row rod-end bearing. Random fusion data augmentation was used for the first 1000 images, with one image expanded and augmented to 12 images to create a dataset containing 13,110 images. The training, validation, and test sets contained 8888, 2222, and 2000 images, respectively. To validate the proposed algorithm, the T-LESS dataset was selected for target detection, because the nature of the data is similar to that for the actual project. Therefore, this dataset was used to evaluate the algorithm performance. The COCO dataset was also used for comparative experiments. Fig. 9b shows sample images from the rod-end bearing dataset (RJB dataset), and Fig. 9c shows sample images from the T-LESS dataset.

The training was performed using a modified YOLOv5 model. We set both the learning rate and recurrent learning rate to 0.01, the stochastic gradient descent momentum to 0.937, the weight decay coefficient to 0.0005, and the batch size to 8. The loss-function curves of the model training and validation are shown in Fig. 10.

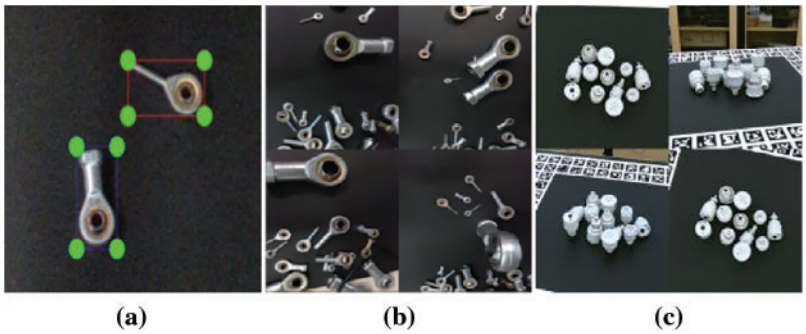


**Figure 9:** Experimental dataset. (a) Annotation dataset; (b) RJB dataset; (c) T-LESS dataset

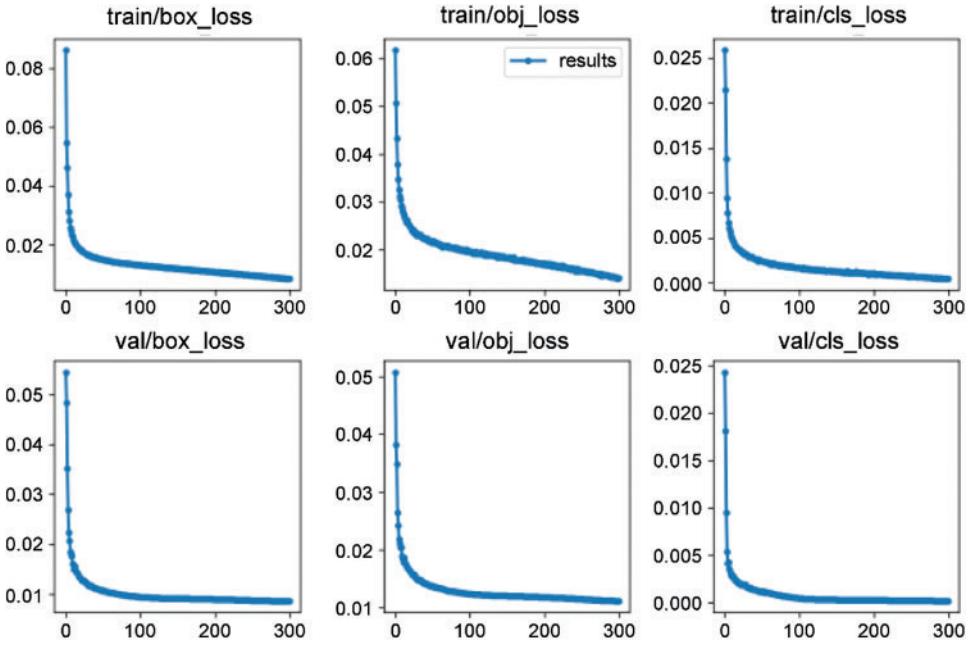


**Figure 10:** Curves of various loss functions during the training process

Here, the vertical coordinate represents the ratio of iterations to epochs, and the horizontal coordinate represents the loss, where box_loss, obj_loss, and cls_loss represent the means of the prediction box GIoU Loss, target detection loss, and label classification, respectively, and "train" and "val" denote the training and validation stages, respectively. Because the rod-end joint bearings were divided into two label categories, i.e., inner tooth rod-end joint bearing SI and outer tooth rod-end joint bearing SA, the loss converged rapidly. As the weighted k-means algorithm was used to obtain suitable anchor boxes, the prediction box and target detection loss converged rapidly, before 100 iterations. The weighted k-means algorithm significantly affects the SCP-YOLOv5 algorithm model.

## 3 Results and Discussion

### *3.1 Analysis of Evaluation Indices and Results*

The experimental indices were the precision (P), recall (R), mean average precision (mAP), and each AP category (mAP). The corresponding formulas are as follows:

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{AllDetections}, \tag{9}$$

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{AllGroundTruths}, \tag{10}$$

$$AP = \int_0^1 PRdr, \tag{11}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP. \tag{12}$$

An ablation experiment was conducted to verify the effectiveness and generalization of the improved algorithm with the addition of the weighted k-means algorithm, improved CP-Cluster algorithm, SPD-Conv module, and MFSPD module (only in the neck)—corresponding to Schemes 1 to 4 in Table 2—and each model was compared with the final improved algorithm SCP-YOLOv5.

**Table 2:** Results of the ablation experiment

| Algoritm | Weighted k-means | Improved CP-Cluster | SPD-Conv | MFSPD | P (%) | | R (%) | | mAP@0.5 (%) | mAP@0.5:0.95 (%) | FPS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | SI | SA | SI | SA | | | |
| YOLOv5s | × | × | × | × | 94.6 | 93.8 | 90.4 | 90.1 | 93.7 | 89.5 | 121 |
| Scheme 1 | √ | × | × | × | 95.3 | 94.2 | 91.8 | 90.8 | 94.2 | 90.1 | 113 |
| Scheme 2 | × | √ | × | × | 95.1 | 94.1 | 90.7 | 90.5 | 94.4 | 89.8 | 119 |
| Scheme 3 | × | × | √ | × | 96.7 | 96.5 | 93.6 | 93.4 | 94.9 | 91.6 | 102 |
| Scheme 4 | × | × | × | √ | 95.2 | 94.3 | 90.9 | 91.3 | 94.5 | 90.8 | 109 |
| SCP YOLOv5 | √ | √ | √ | √ | 98.8 | 98.5 | 96.3 | 96.3 | 96.9 | 92.5 | 106 |

The experimental data in Table 2 indicate the following:

(1) The mAP of the SCP-YOLOv5 algorithm reached 96.9% with a threshold of 0.5, and the detection accuracy reached 92.5% with a threshold between 0.5 and 0.95.

(2) In Scheme 1, the addition of the weighted k-means algorithm increased the precision for SI and SA by 0.7% and 0.4%, respectively, indicating an improvement in the matching between the prior boxes and feature map layers in the model.

(3) In Scheme 2, with the addition of the improved CP-Cluster to the YOLOv5s model, all the metrics increased slightly, and the FPS increased, indicating the efficacy of the parallel processing and DIoU metrics of the improved algorithm.

(4) The detection accuracy for SA was slightly lower than that for SI because there were fewer SA samples in the dataset. However, in Scheme 3, adding the SPD-Conv module to the detection model increased the precision for SI and SA by 2.1% and 2.7%, respectively. Meanwhile, the recall for SI and SA increased by 3.2% and 3.3%, respectively. This indicates that the SPD-Conv module is effective for small-object detection, improving the information extraction capabilities of the network and increasing its detection accuracy.

(5) In Scheme 4, adding the MFSPD module to the neck improved all the metrics, indicating a significant enhancement in the feature fusion capability of the model neck.

The ablation experiment indicated that the weighted k-means, improved CP-Cluster, SPD-Conv, and MFSPD modules positively affect the model. Additionally, the combination of all four achieved the best results, with the precisions for SI and SA reaching 98.8% and 98.5%, respectively.

### *3.2 Comparison with Different Algorithms*

To further verify the performance and advantages of the proposed algorithm, it was compared with mainstream algorithms, i.e., YOLOv3–YOLOv7, for the rod-end bearing dataset, as shown in Table 3.

**Table 3:** Detection performance of different algorithms

| Algorithm | RJB dataset | T-LESS | FLOPs (G) | Number of parameters (M) |
|---|---|---|---|---|
| | mAP@0.5 (%) | | | |
| YOLOv3 | 88.6 | 80.3 | 140.7 | 61.5 |
| YOLOv3-tiny | 85.3 | 76.6 | 5.6 | 8.2 |
| YOLOv4 | 90.7 | 87.3 | 119.8 | 52.5 |
| YOLOv4-tiny | 87.2 | 85.4 | 6.9 | 5.9 |
| YOLOv5s | 93.7 | 91.8 | 16.1 | 7.1 |
| Proposed | 96.9 | 93.8 | 25.3 | 9.5 |
| YOLOv6s | 93.5 | 91.7 | 43.9 | 17.3 |
| YOLOv7 | 98.7 | 95.3 | 103.4 | 36.3 |
| YOLOv7-tiny | 94.8 | 92.2 | 13.2 | 6.1 |

The target and background contrasts of the RJB dataset are evident; therefore, the overall effect is better than that of the T-LESS dataset. The mAP values of YOLOv3 and YOLOv4 were 80.3% and 87.3% respectively at a threshold above 0.5, which were lower than that of the proposed algorithm. Additionally, the large numbers of model parameters and FLOPs make deployment on mobile devices difficult. YOLOv3-tiny and YOLOv4-tiny had fewer parameters and FLOPs than the proposed algorithm. Their detection speeds were higher, but their accuracies were lower. Compared with YOLOv6 and YOLOv7, the mAP@0.5 values of the algorithm were 2.1% higher and 1.5% lower, respectively; however, the number of parameters and FLOPs exceeded those of the proposed algorithm.

As shown in Table 4, the proposed algorithm improves the mAP on the COCO dataset by 1.3% compared to the previous version. This approaches the performance of YOLOv7 and surpasses other mainstream algorithms. Therefore, SCP-YOLOv5 demonstrates balanced performance compared to popular existing methods, with competitive strengths.

To further investigate the significance of differences between the four models (YOLOv5s, SCP-YOLOv5, YOLOv6s, and YOLOv7), we evaluated their performance using Repeated Measures ANOVA and a Friedman Test based on mAP@0.5 results from 10 experimental trials for each model,as shown in Table 5. The Repeated Measures ANOVA yielded a statistically significant result ($F(3, 27) = 252.38$, $p < 0.0001$), indicating significant performance differences among the models. Additionally, the non-parametric Friedman Test confirmed these findings, with a $x^2$ statistic of 28.12 and $p < 0.0001$, further supporting the conclusion that the models differ significantly in terms of detection accuracy. These results suggest that SCP-YOLOv5 and YOLOv7 significantly outperform

YOLOv5s and YOLOv6s in small object detection tasks, validating the effectiveness of the proposed improvements in SCP-YOLOv5.

**Table 4:** Performance of different algorithms for the COCO dataset

| Algorithm | mAP (%) | FPS | GPU |
|---|---|---|---|
| Faster R-CNN [5] | 34.7 | 7 | RTX3050Ti |
| Mask R-CNN [6] | 37.8 | 5 | RTX2070 |
| YOLOv3 | 33.6 | 35 | RTX3050Ti |
| YOLOv4 | 43.5 | 62 | RTX2070 |
| YOLOv5s | 54.6 | 126 | RTX3050Ti |
| YOLOv6 | 51.9 | 135 | RTX2070 |
| YOLOv7 | 56.4 | 161 | RTX2070 |
| Proposed | 55.9 | 106 | RTX3050Ti |

**Table 5:** The results of 10 tests on the RJB dataset for different models

| Algorithm | mAP@0.5 (%) |
|---|---|
| YOLOv5s | 90.4, 90.2, 90.6, 90.3, 90.5, 89.2, 88.9, 86.7, 89.3, 89.7 |
| SCP-YOLOv5 | 96.9, 96.8, 97.0, 96.0, 96.9, 95.7, 96.7, 96.2, 95.8, 96.4 |
| YOLOv6s | 93.5, 93.6, 93.4, 93.7, 93.5, 93.3, 92.8, 92.7, 93.1, 92.5 |
| YOLOv7 | 98.7, 98.6, 98.8, 98.7, 98.6, 97.3, 97.1, 98.1, 98.2, 97.9 |

### *3.3 Error Analysis and Outlook*

Although the detection accuracy of the SCP-YOLOv5 algorithm for the homemade dataset was increased by 3.1% for mAP@0.5 compared with the original YOLOv5s algorithm, a 3.2% error remained.

The main sources of error may have been the following:

(1) Owing to the light source conditions of the experimental environment, small bearings appeared to be reflective during image capture, resulting in a loss of key feature information.

(2) The dataset was insufficient. The top sliding spherical surface of the rod-end bearing had a highly variable structure with different angles.

(3) Although the deep feature fusion module enhances the feature fusion, room for improvement exists in feature representation capabilities.

In future research, we plan to replace the light source to ensure that the light evenly illuminates the surface of the target object and retains the maximum amount of feature information, increase the number of sliding spherical images from various angles to increase the diversity of the dataset, prune the model, and compress and optimize the model to reduce the number of parameters and the computational load while maintaining its accuracy.

### *3.4 Detection Results*

To intuitively demonstrate the detection effect of the improved algorithm, the YOLOv5s and SCP-YOLOv5 algorithms were used for detection with the rod-end-bearing dataset, as shown in Fig. 11 (YOLOv5 algorithm on the left, SCP-YOLOv5 algorithm on the right).

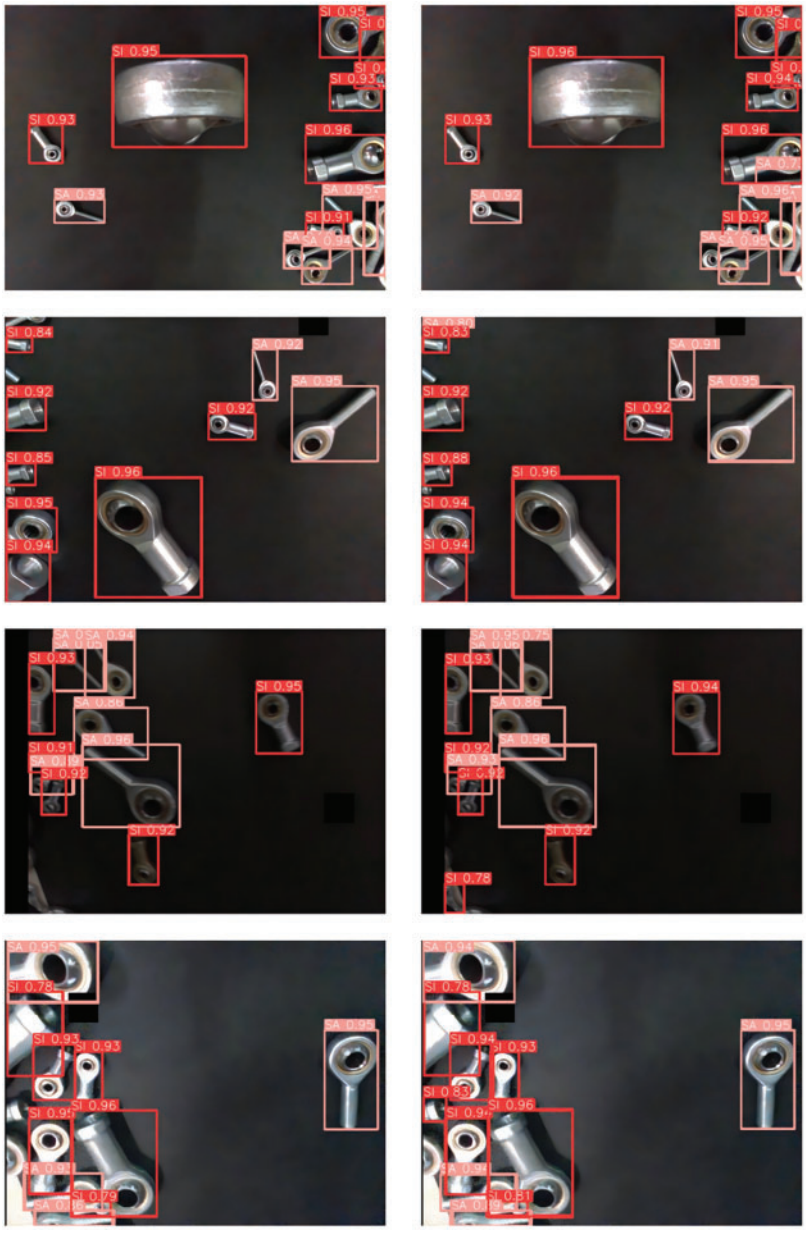

**Figure 11:** Comparison of test results

As shown, the confidence of the SCP-YOLOv5 algorithm for small-bearing detection was higher than that for the YOLOv5 algorithm. Evidently, SCP-YOLOv5 can detect small objects well, and the detection frame is closer to the target, indicating that the improved model is more suitable for practical applications.

## 4 Conclusions

Detection was performed with rod-end bearing, T-LESS, and COCO datasets using various algorithms. In the experiments, the SCP-YOLOv5 algorithm achieved the most balanced performance and good detection results, meeting the requirements of practical applications, with regard to both detection accuracy and speed. It has a significantly higher capability to extract feature information from small rod-end bearings. Compared with the YOLOv5s algorithm, the SCP-YOLOv5 algorithm improved the mAP@0.5 by 3.2% and 2.0% for the RJB and T-LESS datasets, respectively. Furthermore, it improved the mAP by 1.3% for the COCO dataset. This study provides new research perspectives on overcoming the limitations of insufficient feature information and weak contextual representation in small-object detection. It also proposes a novel solution for detecting randomly stacked small parts in CNC machine tool production environments. Although the model detection accuracy was improved, the detection speed was lower than that before the improvements. Future studies may aim to increase the detection speed while ensuring detection accuracy.

**Acknowledgement:** None.

**Funding Statement:** This study was supported by the Fuxiaquan National Independent Innovation Demonstration Zone High End Flexible Intelligent Packaging Equipment Collaborative Innovation Platform Project (2023-P-006); the Program for Innovative Research Team in Science and Technology at Fujian Province University (grant number MinJiaoKe [2020] No.12); and the 2022 Fujian Province Key Scientific and Technological Innovation Project (grant number 2022G02007). The funders had no role in the study design; in the collection, analysis or interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

**Author Contributions:** Jinmin Peng drafted the manuscript and developed the methodology. Ruifeng Ye conducted the investigation, designed and trained the models, and edited the manuscript. Song Lan contributed to data collection and processing. Tenghao Xiao contributed to the formal analysis. Chen Xu contributed to conceptualization. Yancong Song contributed to project administration. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** All data that support the findings of this study are included within the article.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

1. Y. F. Liu, H. L. Qin, C. H. Han, J. D. Shi, G. H. Ma and H. D. Wang, "Current status of life test and damage failure mechanism of self-lubricated joint bearings," (in Chinese), *Mater. Guide*, vol. 35, no. 6, pp. 1036–1045, 2021. doi: 10.11896/cldb.19110135.
2. Y. Cao, H. Li, and T. Wang, "A review of research on target detection algorithms based on deep learning," (in Chinese), *J. Comput. Mod.*, vol. 5, no. 2, pp. 63–69, 2020. doi: 10.3969/j.issn.1006-2475.2020.05.011.
3. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 580–587. doi: 10.1109/CVPR.2014.81.

4. R. Girshick, "Fast R-CNN," in *2015 IEEE Conf. Comput. Vis*, Santiago, Chile, 2015, pp. 1440–1448. doi: 10.1109/ICCV.2015.169.
5. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.
6. K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 386–397, 2020. doi: 10.1109/TPAMI.2018.2844175.
7. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 779–788.
8. J. Terven, D. M. Córdova-Esparza, and J. A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach Learn. Knowl. Extr.*, vol. 5, no. 3, pp. 1680–1716, 2023. doi: 10.3390/make5040083.
9. W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany, 2016, pp. 21–37.
10. Y. Q. Zhang, R. Yuan, S. P. Deng, and J. Y. Zhang, "A review of deep learning target detection methods," (in Chinese). *Chin. J. Image Graph*, vol. 25, pp. 629–654, 2020.
11. D. Kim, J. Seo, S. Noh, and J. Lee, "Compatibility review for object detection enhancement through super-resolution," *Sensors*, vol. 24, no. 11, 2024, Art. no. 3335. doi: 10.3390/s24113335.
12. D. Wahyudi, I. Soesanti, and H. A. Nugroho, "Toward detection of small objects using deep learning methods: A review," in *Proc Int. Conf. Inf. Technol. Electr. Eng. (ICITEE)*, Yogyakarta, Indonesia, 2022, pp. 314–319. doi: 10.1109/ICITEE56407.2022.9954101.
13. C. W. Park, Y. Seo, T. J. Sun, G. Lee, and E. N. Huh, "Small object detection technology using multi-modal data based on deep learning," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, IEEE, 2023. doi: 10.1109/ICOIN56518.2023.10049014.
14. H. Lian and R. Yu, "Progress in small object detection based on deep learning," *J. Aviat.*, vol. 42, pp. 107–125, 2021.
15. M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," in *Int. Conf. Neural Inf. Process.*, Springer, 2021, pp. 387–395.
16. R. Jiang, Y. Peng, W. Xie, and G. Xie, "Improved YOLOv4 small object detection algorithm embedded in the scSE module," *J. Atlas Press*, vol. 42, no. 4, pp. 546–555, 2021.
17. R. Gai, N. Chen, and H. Yuan, "A detection algorithm for cherry fruits based on the improved YOLO-v4 model," *Neural Comput. Appl.*, vol. 35, pp. 13895–13906, 2023.
18. J. Hu, C. J. R. Shi, and J. Zhang, "Saliency-based YOLO for single target detection," *Knowl. Inf. Syst.*, vol. 63, pp. 717–732, 2021. doi: 10.1007/s10115-020-01538-0.
19. J. H. Luo, J. Huang, and X. Y. Bai, "Road small target detection method based on improved YOLOv3," (in Chinese), *J. Chin. Comput. Syst.*, vol. 43, no. 3, pp. 449–455, 2022.
20. X. Wang, H. Li, X. Yue, and L. Meng, "A comprehensive survey on object detection YOLO," in *Proc. 2023 Int. Conf. Artif. Intell. Comput. Vis. (AICV 2023)*, Kusatsu, Japan, 2023, pp. 1–10.
21. C. Li *et al.*, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.
22. C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
23. A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
24. D. Xue, W. Lu, and L. Fan, "Review of the study of typical object detection algorithms for deep learning," *Comput. Eng. Appl.*, vol. 57, no. 8, pp. 10–25, 2021.

25. R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," *Mach. Learn. Knowl. Discov. Databases*, pp. 443–459, 2023. doi: 10.1007/978-3-031-25339-7_22 .
26. W. G. Li, X. Ye, Y. T. Zhao, and W. B. Wang, "Detection of strip steel surface defects based on an improved YOLOv3 algorithm," (in Chinese), *Electron. J.*, vol. 48, no. 7, pp. 1284–1292, 2020.
27. Y. Shen, W. Jiang, Z. Xu, R. Li, and J. Kwon, "Confidence propagation cluster: Unleash full potential of object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 1151–1161.
28. M. S. M. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6626–6634.
29. Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 12993–13000, 2020. doi: 10.1609/aaai.v34i07.6999 .