electronics

Article

ESD-YOLOv5: A Full-Surface Defect Detection Network for Bearing Collars

Jiale Li [1,2], Haipeng Pan [1,2,*] and Junfeng Li [1,2]

[1] School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China; 202130605217@mails.zstu.edu.cn (J.L.); ljf2003@zstu.edu.cn (J.L.)
[2] Changshan Research Institute, Zhejiang Sci-Tech University, Quzhou 324299, China
* Correspondence: pan@zstu.edu.cn

Abstract: To address the different forms and sizes of bearing collar surface defects, uneven distribution of defect positions, and complex backgrounds, we propose ESD-YOLOv5, an improved algorithm for bearing collar full-surface defect detection. First, a hybrid attention module, ECCA, was constructed by combining an efficient channel attention (ECA) mechanism and a coordinate attention (CA) mechanism, which was introduced into the YOLOv5 backbone network to enhance the localization ability of object features by the network. Second, the original neck was replaced by the constructed Slim-neck, which reduces the model's parameters and computational complexity without sacrificing accuracy for object detection. Furthermore, the original head was replaced by the decoupled head from YOLOX, which separates the classification and regression tasks for object detection. Last, we constructed a dataset of defective bearing collars using images collected from industrial sites and conducted extensive experiments. The results demonstrate that our proposed ESD-YOLOv5 detection model achieved an mAP of 98.6% on our self-built dataset, which is a 2.3% improvement over the YOLOv5 base model. Moreover, it outperformed mainstream one-stage object detection algorithms. Additionally, the bearing collar surface defect detection system developed based on our proposed method has been successfully applied in the industrial domain for bearing collar inspection.

Keywords: convolutional neural network; ESD-YOLOv5; bearing collar; defect detection

Citation: Li, J.; Pan, H.; Li, J. ESD-YOLOv5: A Full-Surface Defect Detection Network for Bearing Collars. *Electronics* **2023**, *12*, 3446. https://doi.org/10.3390/electronics12163446

Academic Editors: Haibin Wu, Aili Wang and Yuji Iwahori

Received: 12 July 2023
Revised: 1 August 2023
Accepted: 2 August 2023
Published: 15 August 2023

1. Introduction

Bearings are an important component in mechanical equipment that mainly support the rotation of mechanical components, reduce the friction coefficient during movement, and ensure the accuracy of rotation. The quality of bearings will significantly affect the stability of equipment operation. In the production process, bearings are inevitably affected by factors such as raw materials, processing technology, processing equipment, and external conditions, leading to defects. These defects can result in reduced service life of the bearings and even mechanical equipment failure. Therefore, it is necessary to conduct quality inspections on bearings before they leave the factory.

Currently, defect detection methods are mainly divided into traditional machine vision detection methods and deep-learning-based detection methods. Traditional machine vision detection methods rely on manually extracting defect features and require designing corresponding detection methods for different types of bearing defects. However, bearing defects are diverse in terms of their types, sizes, shapes, and positions, and therefore, manually extracted features cannot adapt to all defects. Deep-learning-based detection algorithms have strong feature expression ability, generalization ability, and cross-scene ability and thus have been widely applied in the industrial field for defect detection. Examples of such applications include detecting features of textiles [1], light guide plates [2], wire and arc additive manufacturing [3], wind turbine gearbox gears [4], and road damage [5].

The YOLOv5 [6] network is currently one of the most commonly used object detection frameworks. It builds upon the foundation of YOLOv4 [7] and introduces several enhancements such as the SPPF (Spatial Pyramid Pooling Fast) module, the CIOU (Complete Intersection over Union) loss function, and adaptive anchor boxes. These advancements contribute to improved detection accuracy and efficiency. YOLOv6 [8] and YOLOv7 [9], on the other hand, focus more on efficiency improvements. Considering the overall performance, we have selected YOLOv5 as the most suitable choice for defect detection in bearing collars. Its combination of improved detection accuracy and efficiency aligns well with the requirements of our study. However, direct use of the YOLOv5 algorithm to identify bearing collar defects does not yield satisfactory results, mainly because bearing collar images have complex backgrounds and a wide variety of defect types, shapes, and sizes. Based on the surface optical characteristics and imaging features of bearing collar defects, as well as the requirements of industrial inspection, we propose an improved YOLOv5-based algorithm for detecting surface defects on bearing collars. Based on the three proposed improvements (ECCA, Slim-neck, and Decoupled head), we have named this model ESD-YOLOv5. Additionally, a detection system for surface defects on bearing collars was developed. The primary contributions of this study are listed as follows:

(1) A hybrid attention mechanism ECCA module was constructed by combining the efficient channel attention mechanism (ECA) [10] and coordinate attention mechanism (CA) [11], which was integrated into the backbone network of YOLOv5 to enhance the feature extraction capability of the network.
(2) The Slim-neck module [12], which combines GSConv and VoVGSCSP, was proposed to replace the Conv and C3 modules in the neck network of YOLOv5. This can effectively reduce the number of parameters while improving the detection capability for defects.
(3) The decoupled head from YOLOX [13] was utilized to replace the original head in order to separate the regression and classification tasks and improve the network's ability to distinguish among the defect categories.

With these three improvements, the ESD-YOLOv5 model achieved an mAP of 98.6% on our custom dataset, which is a 2.3% improvement compared to the original YOLOv5 model. Furthermore, the ESD-YOLOv5 model demonstrated superior performance compared to other mainstream one-stage object detection algorithms. In our work, the ESD-YOLOv5 model exhibited high detection accuracy, precise classification, and a low omission rate, making it highly effective for conducting detection tasks.

The paper is organized as follows: Section 2 reviews the related work; Section 3 presents the composition of the bearing collar defect detection system; Section 4 introduces the network structure of the detection algorithm; Section 5 describes the dataset and experiments; and Section 6 concludes the work presented in this paper.

2. Related Work

2.1. Object Detection Algorithms

Object detection algorithms are divided into one-stage algorithms and two-stage algorithms. Two-stage algorithms generate prediction boxes and then return the location and category information of the object in the prediction box. Representative algorithms include RCNN [14], Fast-RCNN [15], Faster-RCNN [16], etc. One-stage algorithms directly return the position and class information of the targets without generating prediction boxes. Representative algorithms include SSD [17] and YOLO series [6–9,13,18–21]. Generally, two-stage detection algorithms have higher accuracy than one-stage algorithms. However, their detection speed is slower, while real-time detection is usually required in industrial settings. Therefore, one-stage algorithms are more widely employed in industry.

Typically, an object detection network consists of three main components: the backbone, neck, and head. The backbone is responsible for feature extraction, while the neck fuses the features extracted by the backbone at different scales. The head is responsible for predicting the location and category information of the objects. Commonly employed

backbones include VGG [22], ResNet [23], and DarkNet [20], which are based on standard convolutions and typically have many parameters and computational requirements. To address this issue, lightweight backbones, such as MobileNet [24–26], ShuffleNet [27,28], and GhostNet [29,30], have been proposed. For the neck, there are two main structures for feature fusion and enhancement: the feature pyramid network (FPN) [31] and the path aggregation network (PAN) [32]. The choice of head depends on whether the model uses anchor-based or anchor-free methods for object detection. The former generally achieves higher accuracy, while the latter is more flexible. In addition, attention mechanism modules, such as SE [33], CBAM [34], ECA [10], and CA [11], can be incorporated to enhance the performance of the network. Moreover, some semi-supervised learning and unsupervised learning methods such as Consistent Teacher [35], Efficient Teacher [36], and MGLNN [37] have also been of great assistance in the field of computer vision.

2.2. Bearing Collar Defect Detection

In recent years, deep learning has gained widespread adoption across various industrial domains. However, there remains a limited body of research on detecting surface defects of bearings using deep learning techniques. For instance, Zheng et al. [38] proposed a bearing cap defect detection method based on an improved YOLOv3 algorithm. This method incorporates attention mechanisms, multiscale feature fusion, anchor box clustering, and other techniques to enhance the detection performance and robustness of bearing cap defects. The experimental results showed that the proposed method achieved an mAP of 69.74%, which is 16.31%, 13.4%, 13%, 10.9%, and 7.2% more than that of YOLOv3, EfficientDet-D2, YOLOv5, YOLOv4, and PP-YOLO, respectively. However, the confidence level for certain target categories in this method still requires improvement. Lei et al. [39] proposed a segmented embedding rapid defect detection method (SERDD) for bearing surface defects. This method achieved bidirectional fusion of image processing and defect detection, resulting in an accuracy of 81.13% for bearing surface character detection and 100% accuracy for bearing surface defect detection. Nonetheless, this method is only effective for a single type of bearing, and further optimization is needed. Xu et al. [40] proposed an unsupervised neural network based on autoencoder networks, which use U-net to create an automatic encoder network for predicting outputs. Compared with the supervised ResNet, this method performed better in detecting defects with limited training samples. The experimental results showed that the method achieved an AUC of 96.23%, outperforming ResNet's 85.67%. However, since the unsupervised neural network is based on the autoencoder network and uses the gradient of unannotated data as labels, it may introduce noise or inaccurate information. Liu et al. [41] employed two lighting modes (coaxial light and multisource light) to capture images of bearings, processed the images using traditional algorithms, and utilized neural networks to detect four common types of defects. The experimental results showed a detection accuracy of 98.75% with an average time consumption of detection of 2.11 s/bearing. However, there may be more types and forms of defects on the surfaces of bearings, so the generalization ability and robustness of the system need to be improved. Fu et al. [42] proposed a two-stage detection method based on convolutional neural networks (CNNs) and improved the segmentation network using attention and spatial pyramid pooling techniques. The experimental results demonstrated an Intersection over Union (IoU) of 85.81%, which is 2.01% higher than the original model. However, the speed of the two-stage detection method was slower.

Although the aforementioned methods have achieved a certain degree of automation and intelligence in bearing surface defect detection, there still exist some gaps and challenges, such as: (1) the lack of large-scale, diverse, and high-quality bearing surface defect image datasets, leading to issues of insufficient, imbalanced, and non-representative training data; (2) the absence of a universal bearing surface defect detection algorithm, resulting in the problem of algorithm instability under different types of defects, working conditions, and lighting conditions; and (3) the lack of efficient and practical bearing surface defect detection systems, leading to limitations in real-time capability and accuracy,

which cannot meet the demands of industrial production. Therefore, in the future, the field of bearing surface defect detection requires in-depth research and innovation from three aspects, data, algorithms, and systems, to elevate the level and application value of bearing surface defect detection.

3. Bearing Collar Defect Detection System

3.1. Bearing Collar Defect Detection Device

The bearing collar defect visual detection device that was designed and developed in this study is shown in Figure 1. The device mainly consists of three parts: a mechanical transmission system, an image acquisition system, and an image processing system. The mechanical transmission system mainly consists of a frame, clamp, and cylinder, to achieve the movement and flipping of the bearing. The image acquisition system consists of three area scan cameras, one line scan camera, and multiple angled light sources, which capture images of the bearing and its defects. The image processing system consists of an industrial computer, detection system software, and other components to achieve accurate and real-time detection of various defects of the bearing collar.

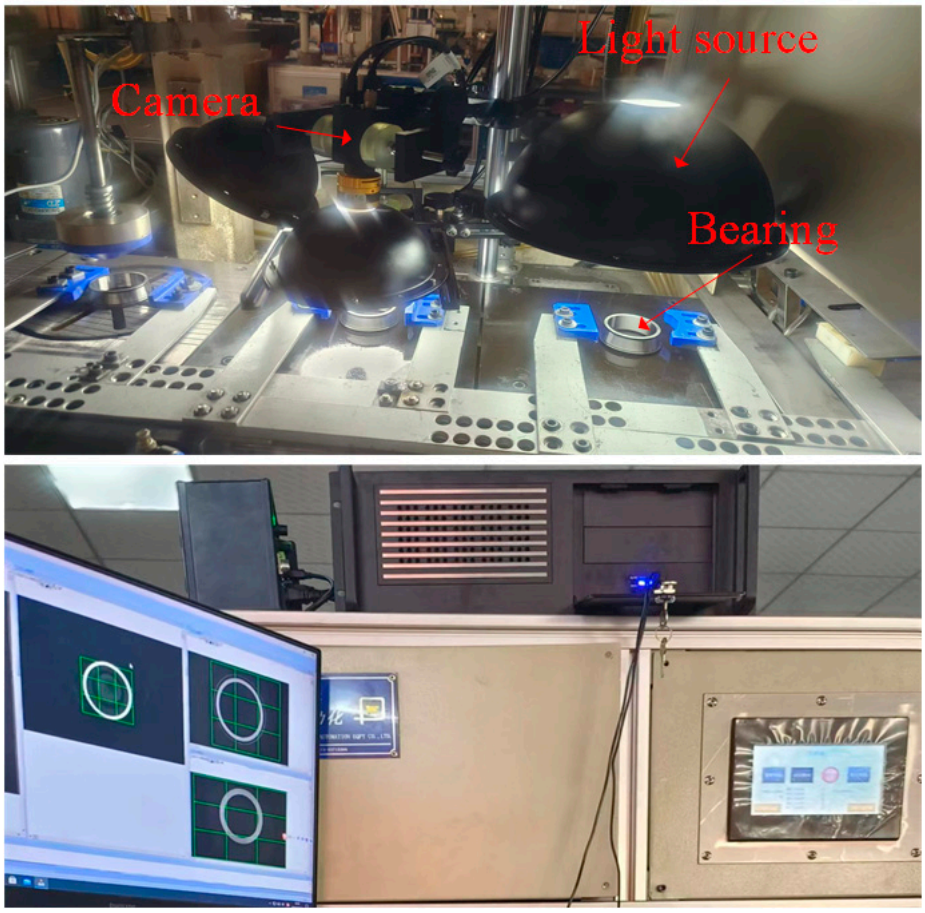
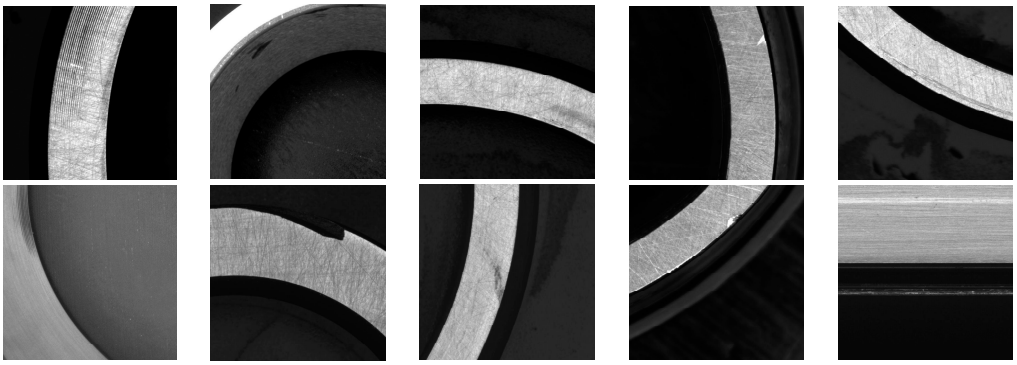


Figure 1. Bearing collar defect detection device.

3.2. Bearing Collar Defects Imaging Analysis

In this study, the image resolution of the area scan cameras was 5472 × 3648 and that of the line scan camera was 2048 × 10,000. The network's detection ability greatly decreases with excessively high resolutions. Therefore, a sliding window with a size of 640 × 640 and a stride of 0.85 was applied to crop the original image into small images for training and detection. As shown in Figure 2, bearing collar defects can be roughly divided into thread, black spot, wear, dent, and scratch defects.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 2. Bearing collar images of different types of defects. (**a**) Thread; (**b**) black spot; (**c**) wear; (**d**) dent; (**e**) scratch defects.

3.2.1. Bearing Collar Defect Imaging Features

(1) Thread

Thread defects, as shown in Figure 2a, are mainly caused by equipment failure or improper bearing collar placement during the lathe machining process. These defects usually appear on the end face and inner side of the bearing collar, manifesting as dense black curves with prominent features.

(2) Black spot

Black spot defects, as shown in Figure 2b, are mainly caused by missing material or rust during the bearing collar forging process. These defects appear on all four surfaces of the bearing collar, with varying sizes and shapes, and are easily confused with the black background.

(3) Wear

Wear defects, as shown in Figure 2c, are mainly caused by the reduction in the bearing collar surface gloss due to friction. They appear on the end face and outer side of the bearing collar and vary greatly in size, shape, and color.

(4) Dent

Dent defects, as shown in Figure 2d, are dents at the edges of the bearing collar, typically appearing on the end face and with relatively small dimensions.

(5) Scratch

Scratch defects, as shown in Figure 2e, are mainly caused by the improper installation of the bearing collar, which leads to collisions between the bearing and other objects. These defects usually appear on the end face and the outer side of the bearing collar, and their sizes and shapes vary. Scratch defects are relatively shallow, but their longitudinal extent can be longer than that of other types of defects.

3.2.2. Difficulties of Bearing Collar Defect Detection

Based on the imaging characteristics and detection requirements of bearing collar defects, there are several main challenges in defect detection:

(1) As the bearing collar is ring-shaped, in this paper, sample images were obtained using a sliding window approach, which produced a somewhat complex background.
(2) Dust and oil stains can appear on the surface of the bearing collar, and their imaging characteristics are very similar to those of defects, which can easily lead to misjudgments.
(3) Black spot defects have the same color as the black background and can only be distinguished by their shape, which can lead to misjudgments.
(4) The sizes of threads, black spots, and wear defects significantly differ, and the detection model needs to simultaneously have a good detection effect on multiscale targets.

4. ESD-YOLOv5

4.1. Network Structure of ESD-YOLOv5

YOLOv5 is an object detection network that is composed of three main components: a backbone, neck, and head. As shown in Figure 3, CSPDarknet53 possesses advantages such as being lightweight, efficiency, and multi-scale adaptability, making it suitable for various object detection tasks in different scenarios. Therefore, we select CSPDarknet53 as the backbone network to extract feature information from input images. As shown in Table 1, the backbone network performs five down-sampling operations on the input image. The down-sampling module CBS consists of convolution, batch normalization, and the SiLU activation function. The C3 module is mainly utilized for feature extraction and is a type of CSP (Cross Stage Partial) structure that is composed of three down-sampling modules (CBS) and multiple bottleneck modules. The SPPF is a spatial pyramid pooling module that performs max pooling with different kernel sizes to increase the network's receptive field and combines the features for fusion. The neck of YOLOv5 adopts an FPN + PAN structure, in which the FPN (feature pyramid network) layer passes strong semantic features from top to bottom, while the PAN (path aggregation network) layer passes strong localization features from bottom to top. Feature aggregation is performed on different detection layers from different backbone layers to enhance the feature extraction capability. The head of YOLOv5, which is a fully convolutional network, was inherited from YOLOv3 and can output three sets of predictions at different scales, each containing the position, confidence, and class of the detected objects. In addition, YOLOv5 can be divided into four versions (YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) based on the model's depth and width. Generally, larger models tend to achieve higher accuracy, but at the expense of a slower speed. In this paper, we have selected YOLOv5s, which is the fastest version, as the base model for improvement. The YOLOv5 backbone network sacrifices some feature extraction ability, resulting in poor performance in detecting small objects. The computation and memory consumption of the neck structure are large, leading to a decrease in the inference speed of the model. The head section needs to simultaneously predict both regression tasks and classification tasks, which can reduce the convergence speed of the loss function. Therefore, we propose three improvements to the YOLOv5 architecture; the improved network structure is shown in Figure 3. The ECCA module is a hybrid attention mechanism proposed in this study that integrates the ECA and CA mechanisms to enable the network to focus more on channel and spatial information of features. The CBS module in the neck structure was replaced with GSConv, and the C3 module was replaced with VoVGSCSP. GSConv has a lower computational cost and produces better results than standard convolution in terms of computation. The head of YOLOv5 is replaced with the decoupled head from YOLOX, which separates the classification and regression tasks and significantly accelerates the convergence of the loss function.

Table 1. The detailed structure of backbone.

Type	Size	Stride	Filters	Output
Convolutional	6 × 6	2	64	320 × 320 × 32
Convolutional	3 × 3	2	128	160 × 160 × 64
C3	-	-	128	160 × 160 × 64
Convolutional	3 × 3	2	256	80 × 80 × 128
C3	-	-	-	80 × 80 × 128
Convolutional	3 × 3	2	512	40 × 40 × 256
C3		-	-	40 × 40 × 256
Convolutional	3 × 3	2	1024	20 × 20 × 512

Table 1. *Cont.*

Type	Size	Stride	Filters	Output
C3	-	-	-	20 × 20 × 512
ECCA	-	-	-	20 × 20 × 512
SPPF	5 × 5	-	1024	20 × 20 × 512

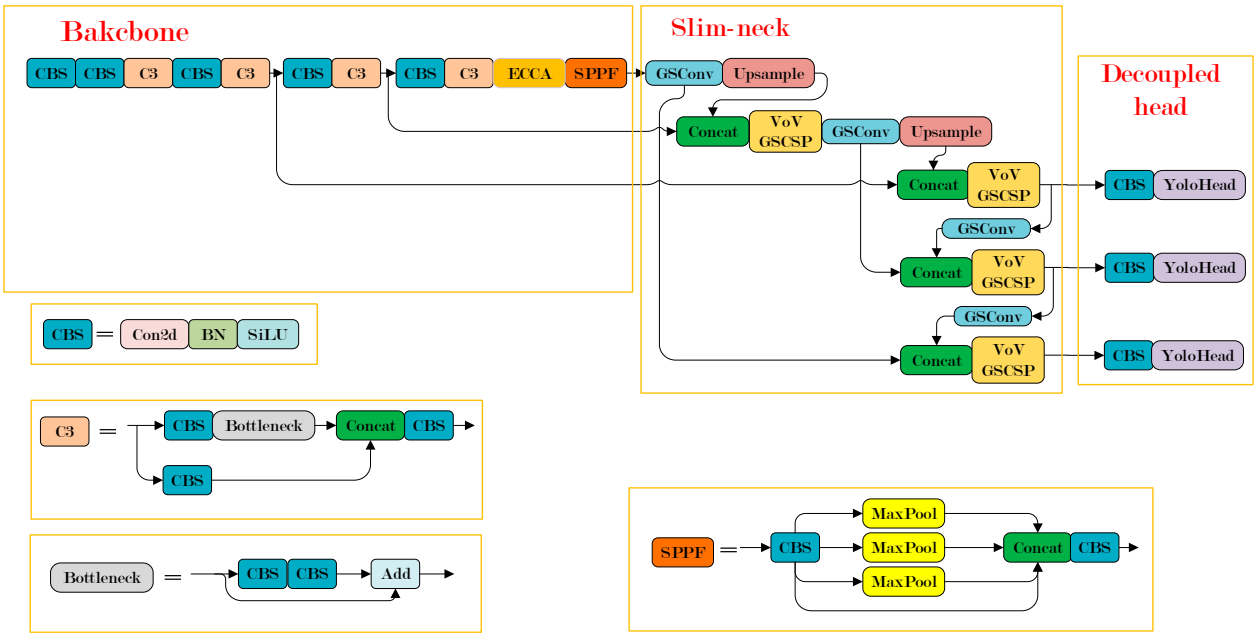


Figure 3. Network structure of ESD-YOLOv5.

4.2. ECCA Module

The attention mechanism is essentially similar to the selective visual attention mechanism of humans. The mechanism adjusts the weights of different regions of an image so that we can focus more on important areas while disregarding irrelevant information. Attention mechanisms have been proven to be effective in various computer vision tasks, such as image classification and object detection. Therefore, incorporating attention mechanisms can enable the network to focus more on defect regions. The function of the CA module is to decompose the channel attention into two 1D-feature-encoding processes, which aggregate features along the H and W spatial directions. This decomposition allows for capturing remote dependency relationships along one spatial direction while preserving accurate position information along the other spatial direction. Then, the generated feature maps are separately encoded into a pair of direction-aware and position-sensitive attention maps, which can be complementarily applied to the input feature map. The CA module takes into account both channel relationships and positional information. The module captures not only channel information but also direction-aware and position-sensitive information, which enables the model to more accurately locate and recognize object areas. However, because the CA module needs to simultaneously consider the channel and positional information of the feature map, training may result in the loss of channel information. As a lightweight channel attention module, the ECA module can capture cross-channel interactions and achieve significant performance improvements. The ECCA module was constructed by combining the CA and ECA modules. The ECA module is utilized to assist in capturing channel information within the CA module, and the resulting ECCA module was introduced into the backbone network of YOLOv5 to achieve better feature extraction.

4.2.1. CA

The CA module is illustrated in Figure 4. First, global average pooling is applied along the horizontal and vertical directions to obtain two separate position-sensitive feature maps, where the result of vertical pooling is permuted to swap the second and third dimensions. Second, the two feature maps are concatenated along the spatial dimension and encoded with Conv, BN, and hardSwish to capture the spatial information in the vertical and horizontal directions. Last, the two position-sensitive feature maps are separated and weighted to be applied to the input feature map.

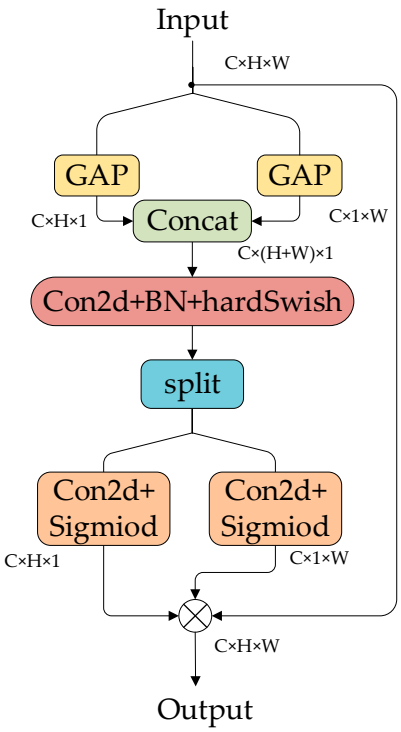


Figure 4. Structure of the CA module.

4.2.2. ECA

The structure of ECA is shown in Figure 5. First, the input feature map (C × H × W) is globally average pooled (G × A × P) to obtain a C × 1 × 1 tensor. Second, fast one-dimensional convolution with a kernel size of k is employed to capture cross-channel interaction information, obtaining the weight values of each channel and generating a C × 1 × 1 feature map through an activation function. Last, the feature map is multiplied elementwise with the input feature map to obtain the final feature map. The ECA module avoids dimensionality reduction by adding only few parameters. To better capture cross-channel interactions, ECA considers each channel and its k adjacent ranges as key indicators. The kernel size k indicates the coverage of the local cross-channel interactions in terms of how many adjacent ranges participate in the attention calculation. The value of k can be adaptively determined based on the number of channels, as shown in Equation (1):

$$k = \phi(c) = \left| \frac{log_2(c)}{r} + \frac{b}{r} \right|_{odd} \tag{1}$$

where c is the number of channel dimensions, $|t|_{odd}$ is the nearest odd number of t, r is set to 2, and b is set to 1.

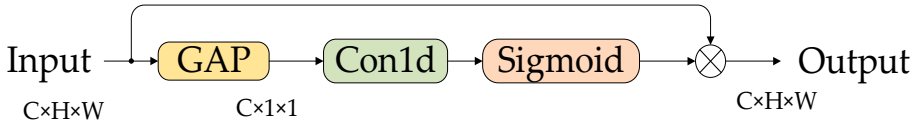


Figure 5. Structure of ECA.

4.2.3. ECCA

The defects of bearing collars are complex and diverse and are often affected by background interference. Some defects cannot be detected using the YOLOv5 model. However, by incorporating attention mechanisms to focus on the features of the defect region, the feature extraction capability of the defect detection network can be improved. In this study, we combined the CA module and ECA module to construct the ECCA module, which we added before the SPPF module in the YOLOV5 backbone network to enhance the network's feature extraction. The structure of ECCA is illustrated in Figure 6.

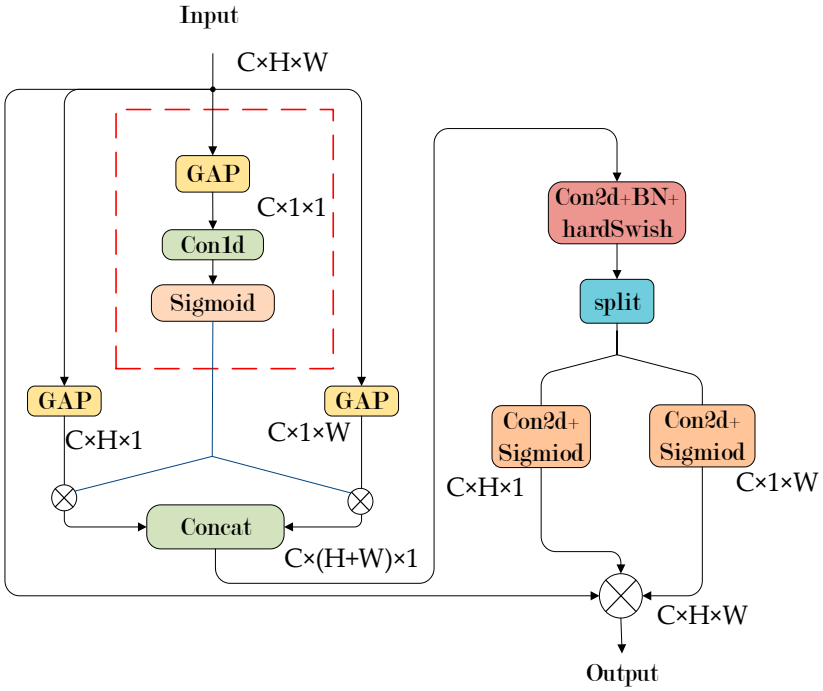


Figure 6. Structure of the ECCA module.

The ECCA module weights the channel feature vectors extracted by ECA and applies them to the two position-aware feature maps of the CA module. These steps are performed to enhance the cross-channel interaction information obtained from the position-aware feature maps and to improve the network's performance, thereby strengthening the feature extraction process.

4.3. Slim-Neck

In industrial projects, detection accuracy and inference requirements are typically high. Usually, the higher the number of parameters of a model is, the higher the detection accuracy. However, the corresponding detection speed may decrease. Therefore, we introduce the lighter convolutional structure GSConv, which can reduce parameters and computation complexity without sacrificing feature expression capability. The GSConv module was embedded in the feature fusion stage to enable the new model to achieve better performance with significantly fewer parameters. We did not use GSConv in the backbone network because it would lead to deeper layers, which would increase the resistance to spatial information flow and affect the inference speed.

4.3.1. GSConv

Figure 7 shows the structure of the GSConv module. The structure of GSConv consists of two parts: a standard convolution (SC) layer and depthwise separable convolution (DSC) [43] layer. The SC layer is responsible for extracting high-level semantic information from the feature map, while the DSC layer reduces the number of channels and computational complexity of the feature map. The feature information extracted by these two layers is then concatenated and passed through a channel shuffle operation to obtain the output feature map. The channel shuffle operation is performed to rearrange channels after grouped convolution, allowing information exchange between different groups and improving the network's performance and accuracy.

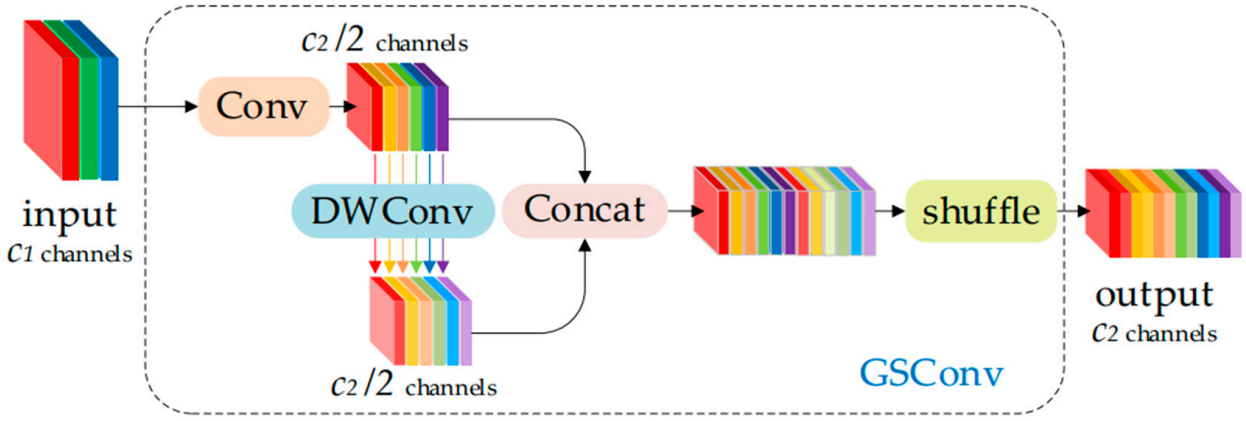


Figure 7. Structure of GSConv module.

The convolutional computation is usually defined by FLOPs (Floating Point Operations). Therefore, the time complexity of SC, DSC, and GSConv is expressed in terms of FLOPs. Specifically, the time complexity of SC, DSC, and GSConv is denoted as follows:

$$Time_{SC} \sim O(W \cdot H \cdot K_1 \cdot K_2 \cdot C_1 \cdot C_2) \tag{2}$$

$$Time_{DSC} \sim O(W \cdot H \cdot K_1 \cdot K_2 \cdot 1 \cdot C_2) \tag{3}$$

$$Time_{GSConv} \sim O\left(W \cdot H \cdot K_1 \cdot K_2 \cdot (C_1 + 1) \cdot \frac{C_2}{2}\right) \tag{4}$$

where W and H represent the width and height, respectively, of the feature map; K_1 and K_2 denote the sizes of the convolutional kernels; and C_1 and C_2 indicate the number of input channels and number of output channels, respectively. These three equations indicate that the time complexity of GSConv is between that of SC and that of DSC.

4.3.2. VoVGSCSP

Based on GSConv, we introduced the GS Bottleneck and VoVGSCSP modules. Figure 8 illustrates the structures of the GS bottleneck and VoVGSCSP modules. Compared with the bottleneck and C3 modules used in YOLOv5, VoVGSCSP reduces the number of parameters and computation by using group convolution and channel shuffling, thus improving the lightweight nature of the model. Furthermore, the model's accuracy is enhanced by increasing the feature extraction capability and receptive field via multibranch convolution.

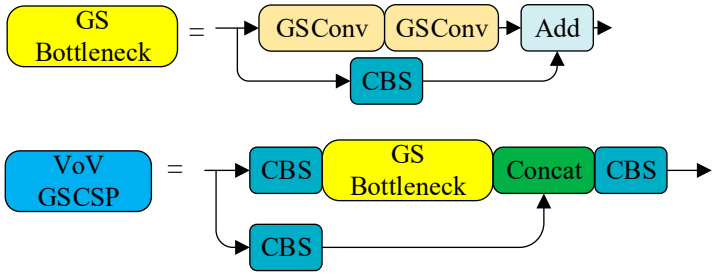


Figure 8. Structure of the GS bottleneck and VoVGSCSP module.

4.3.3. Slim-Neck

The neck of YOLOv5 is a feature fusion network that merges feature maps of three different scales extracted from the backbone network to obtain richer feature information. To balance model accuracy and speed, we used a Slim-neck feature fusion network composed of GSConv and VoVGSCSP. Figure 3 illustrates the network structure of the Slim-neck. In comparison to the neck of YOLOv5, Slim-neck replaces the CBS and C3 modules with GSConv and VoVGSCSP. This replacement enables a reduction in parameters and computational complexity, while simultaneously improving the speed and efficiency of the model.

4.4. Decoupled Head

The head section is the detection part of YOLOv5. In the original YOLOv5 algorithm, a coupled head is utilized, where, after feature fusion, the final detection head is directly obtained by a convolutional layer. The detection head couples position, object, and class information. In contrast, in this paper, we used the YOLOX decoupled head, which is shown in Figure 9. The decoupled head structure consists of a 1×1 convolutional layer that reduces the number of channels, followed by two parallel branches. The first branch is responsible for classification, while the second branch is responsible for regression. The output shape of the classification branch is $H \times W \times C$, and the regression branch is further divided into two branches for position and object confidence, with output shapes of $H \times W \times 4$ and $H \times W \times 1$, respectively. We still used an anchor-based detection mechanism in this study, so each output needs to be multiplied by the number of anchor boxes. As the decoupled head can separately extract classification and regression features, avoiding interference between features, using a decoupled head can greatly accelerate the convergence speed of the loss function during training.

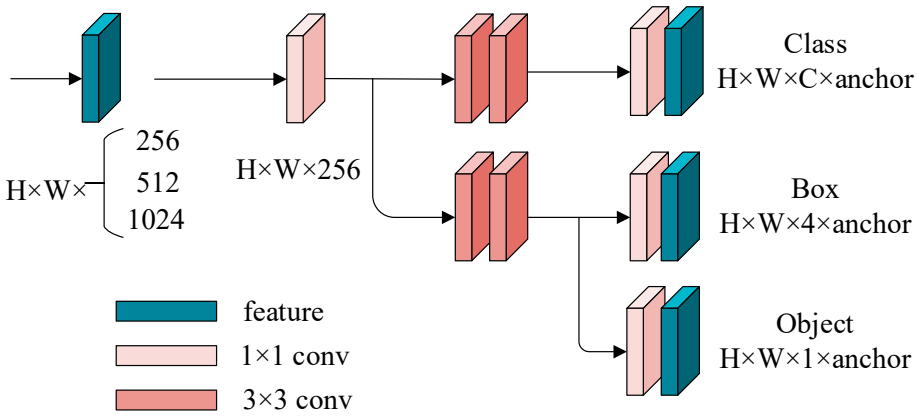


Figure 9. Structure of the decoupled head.

4.5. K-Means Algorithm and Loss Function

4.5.1. K-Means Algorithm

YOLOv5 is an anchor-based object detection algorithm that uses anchor boxes to predict the bounding boxes of objects. The shape and size of the anchor boxes have a

significant impact on the detection performance, so it is necessary to perform clustering analysis based on the characteristics of the dataset to obtain appropriate anchor boxes. The K-means algorithm is an unsupervised clustering algorithm that can divide unlabeled data into a certain number of different groups. The calculation steps are shown in Table 2.

Table 2. K-means calculation steps.

Step 1	K objects are randomly selected from the data as the initial cluster centers.
Step 2	The distance between each data object and the cluster center is computed, and the data object is assigned to the cluster corresponding to the closest cluster center.
Step 3	The mean of data objects in each cluster is calculated to obtain new cluster centers.
Step 4	Steps 2 and 3 are repeated until the cluster centers no longer change or until the maximum number of iterations is reached.

In the anchor calculation of YOLOv5, the bounding boxes are generally considered 2D points (width and height), and the K-means algorithm is used to cluster these points to obtain K anchor boxes that best fit the size of the true boxes. Since YOLOv5 has three different scales of feature maps, each scale of the feature map has three anchor boxes. Thus, we chose to cluster nine anchor boxes. The sizes were (39, 39, 62, 122, 178, 76), (106, 244, 597, 58, 236, 202), and (178, 547, 478, 220, 354, 455), and the anchor boxes of different scales correspond to different sizes of objects.

4.5.2. Loss Function

The loss function is used to measure the degree of closeness between the predicted output of a neural network and the expected output. The smaller the loss function value is, the closer the predicted output is to the expected output. The loss function utilized in YOLOv5 consists of three parts: position loss, object loss, and classification loss. The position loss is applied to measure the distance between the predicted position and the expected position; the object loss represents the probability of the presence of an object, usually a value between 0 and 1, with larger values indicating a higher probability; and the classification loss represents the probability that the object belongs to a certain class. The overall loss function is the weighted sum of the three aforementioned loss functions, as shown in Equation (5):

$$Loss = w_{box}L_{box} + w_{obj}L_{obj} + w_{cls}L_{cls} \tag{5}$$

where w_{box}, w_{obj}, and w_{cls} are 0.05, 0.5, and 1, respectively.

The position loss L_{box} is defined as:

$$L_{box} = 1 - IOU + \frac{\rho^2(A,B)}{c^2} + \alpha\nu \tag{6}$$

where IOU is the intersection over union between the prediction frame and the real frame, and a larger IOU indicates that the real frame is closer to the prediction frame; ρ is the Euclidean distance between the coordinates of the center point of the real box A and the predicted box B; and c is the length of the diagonal of the smallest closed rectangle containing the predicted and ground truth bounding boxes, which is utilized for distance normalization. The weight coefficient α is used to balance the contribution of different loss components, while ν is applied to measure the consistency of the aspect ratio between A and B.

IOU is defined as:

$$IOU = \frac{A \cap B}{A \cup B} \tag{7}$$

where A is the real box, B is the prediction box, $A \cap B$ is the intersection of A and B, and $A \cup B$ is the union of A and B.

α and ν are defined as follows:

$$\alpha = \frac{\nu}{1 - IOU + \nu} \tag{8}$$

$$\nu = \frac{4}{\pi^2}\left(arctan\frac{w^B}{h^B} - arctan\frac{w}{h}\right)^2 \tag{9}$$

In this study, both the object loss and classification loss are calculated using the binary cross-entropy loss function, which is defined as follows:

$$L_{cls} = L_{obj} = -\frac{1}{n}\sum(y_n \times \ln x_n + (1 - y_n) \times \ln(1 - x_n)) \tag{10}$$

where n represents the number of input samples, y_n represents the true value of the target, and x_n represents the predicted value of the network.

5. Experimental Verification

5.1. Bearing Collar Surface Defect Dataset

The bearing collar defect dataset employed in this study was collected from an industrial site, and the bearing collar surfaces that need to be inspected include the upper surface, lower surface, inner surface, and outer surface. The upper, lower, and inner surfaces were imaged using a planar camera with an image resolution of 5472 × 3648. The outer surface was imaged using a linear camera with an image resolution of 2048 × 10,000. A total of 1000 defective bearing collar images were collected and cropped using a sliding window with a size of 640 × 640 and a step size of 0.85. Defect images were then selected, and the dataset was divided into five categories of defects—thread, dark spot, wear, dent, and scratch—based on the features of the defect. Due to the differences in the number of each defect type in actual production, to ensure the rationality of training and balance between each type of defect, the quantity of each defect type was expanded. After expansion, the total number of images was 4934, and the number of labels was 5358. The statistical data for each type of defect after expansion are shown in Table 3. Based on the number of dataset samples and training rationality, the samples of each type of defect were randomly divided into a training set, validation set, and testing set at a ratio of 8:1:1.

Table 3. Expanded defect dataset.

Defect	Thread	Black Spot	Wear	Dent	Scratch	Total
Number	926	1152	1218	812	1250	5358

5.2. Experimental Setting

The hardware environment and software versions for the experiments are shown in Table 4.

Table 4. Experimental environment.

	Configurations
Hardware	Operating system: Ubuntu 18.04 CPU: Intel(R) Xeon(R) Platinum 8358P GPU: RTX A5000
Software	Python: 3.9 CUDA: 11.1 Pytorch: 1.10.0

5.3. Performance Metrics

To verify the effectiveness of the ESD-YOLOV5 defect detection model, this paper applied mean average precision (mAP), parameter quantity, computational complexity

(FLOPs), and frames per second (FPS) as evaluation metrics. "Parameter quantity" refers to the total number of trainable parameters in the model. These parameters are learned during the training process to map input data to output results, including weights and biases, among others. Parameter quantity is an important metric to measure the model's complexity and capacity. Generally, a higher number of parameters indicates a stronger expressive power of the model, but it also means an increase in the computational resources required for training and inference. FPS represents the number of images the object detection network can process per second, and the larger the FPS is, the faster the network processing speed.

The confusion matrix is shown in Table 5.

Table 5. Confusion matrix.

Real \ Prediction	Positive	Negative
True	TP	FN
False	FP	TN

In Table 4, TP (true positive) represents the number of samples that are positive and correctly predicted, FP (false positive) represents the number of samples that are negative but predicted as positive, and FN (false negative) represents the number of samples that are positive but predicted as negative. TN (true negative) represents the number of samples that are negative and correctly predicted.

The precision and recall rates are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

The definitions are presented as follows:

$$AP = \int_0^1 P(R)dR \tag{13}$$

$$mAP = \frac{\sum_{n=0}^{c} AP(C)}{C} \tag{14}$$

where AP is the area under the (P–R curve) formed by the precision and recall and mAP represents the average value of AP for each category, which is used to measure the detection performance of the network model for all categories.

5.4. Ablation Experiments

In this study, we made three improvements to YOLOv5. To verify the effectiveness of each improvement as well as the combination of the three improvements, ablation experiments were conducted. The results are shown in Table 6.

As shown in Table 6, the mAP of YOLOv5s was 96.3%. After adding the ECA module, the mAP increased to 96.7%. Adding the CA module further improved the mAP to 97.0%. The combination of ECA and CA modules in the ECCA module enhanced the network's ability to detect surface defects on bearing collars, resulting in an mAP of 97.8%. When combined with the Slim-neck, the mAP increased to 98.1%, accompanied by a reduction in both the parameters and computational complexity. With the addition of the decoupled head, the highest detection accuracy was achieved with an mAP of 98.6%, indicating a 2.3% improvement over YOLOv5s. However, it should be noted that the Decoupled head significantly increased the parameters and computational complexity, resulting in a decrease in FPS. A total of 269 images were obtained after the cropping process using the

sliding window on the images captured by the four cameras. Theoretically, the detection process can be completed within 3 s using ESD-YOLOv5. However, in industrial settings, the requirement is to complete the detection within 8 s. Therefore, the proposed ESD-YOLOv5 meets the demands of practical bearing production inspections.

Table 6. Results of ablation experiments.

Method	Params (M)	FLOPs (G)	mAP@0.5	FPS
YOLOv5s	7.03	15.8	96.3%	137
YOLOv5s + ECA	7.03	15.8	96.7%	137
YOLOv5s + CA	7.05	16.0	97.0%	135
YOLOv5s + ECCA	7.05	16.0	97.8%	135
YOLOv5s + ECCA + Slim-neck	6.88	14.1	98.1%	148
ESD-YOLOv5	14.20	54.3	98.6%	91

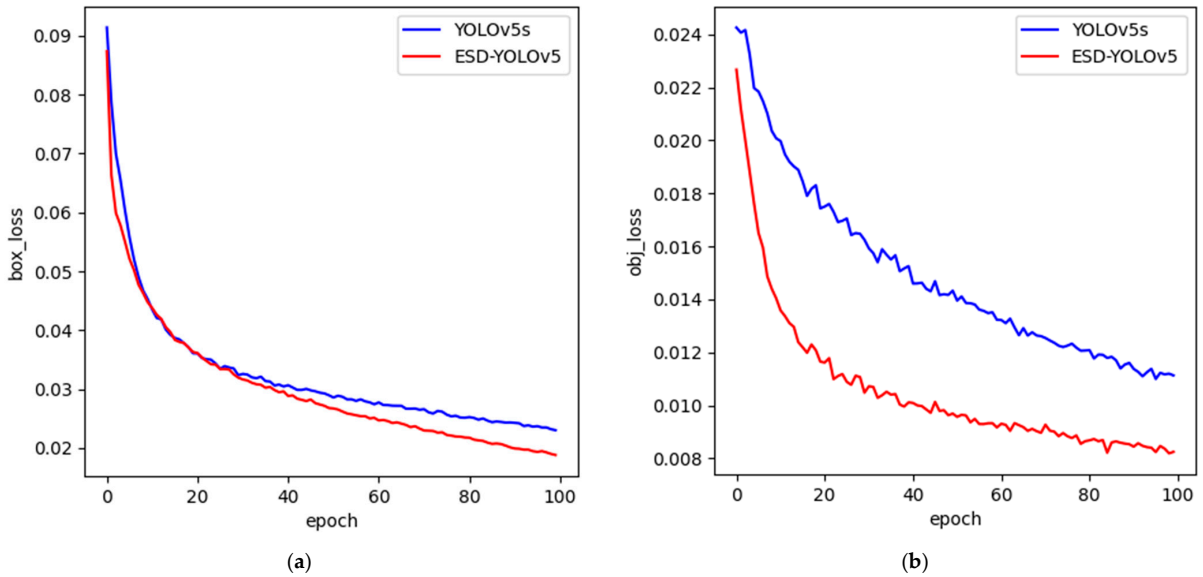
5.5. Comparison Experiments

5.5.1. Experimental Results of Bearing Collar Surface Defect Detection

To further validate the effectiveness of the improved YOLOv5 defect detection model, this study compared it with several other single-stage object detection methods, including YOLOv5, YOLOX, YOLOv6, YOLOv7, and YOLOv8. The training loss and mAP curves during the training process are shown in Figure 10, and the comparison results with the other models are presented in Figure 11. The experimental results are summarized in Table 7.

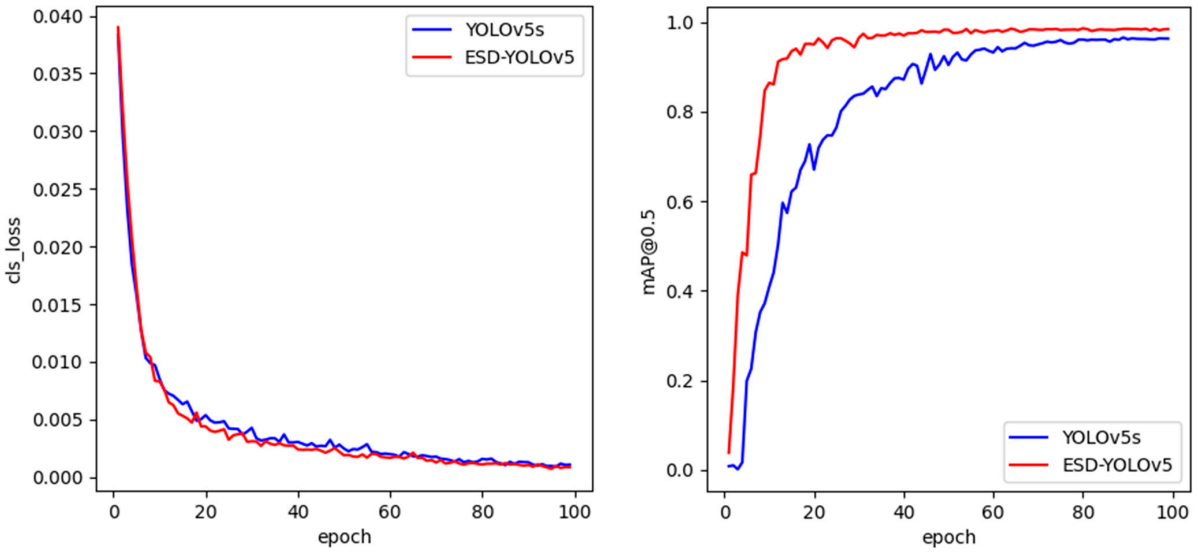
Table 7. Comparison of related methods on bearing collar dataset.

Model	Params (M)	FLOPs (G)	mAP@0.5	FPS
YOLOv5s	7.0	15.8	96.3%	137
YOLOXs	8.7	26.4	95.8%	124
YOLOv6n	4.6	11.3	93.7%	223
YOLOv7tiny	6.0	13.2	94.8%	204
YOLOv8s	11.14	28.7	96.3%	117
YOLOv5m	20.9	48.3	97.5%	96
Ours	14.2	54.3	98.6%	91



(**a**) (**b**)

Figure 10. *Cont.*



(**c**) (**d**)

Figure 10. Training loss and mAP curve of YOLOv5s and ESD-YOLOv5. (**a**) Position loss; (**b**) object loss; (**c**) classification loss; (**d**) mAP@0.5.

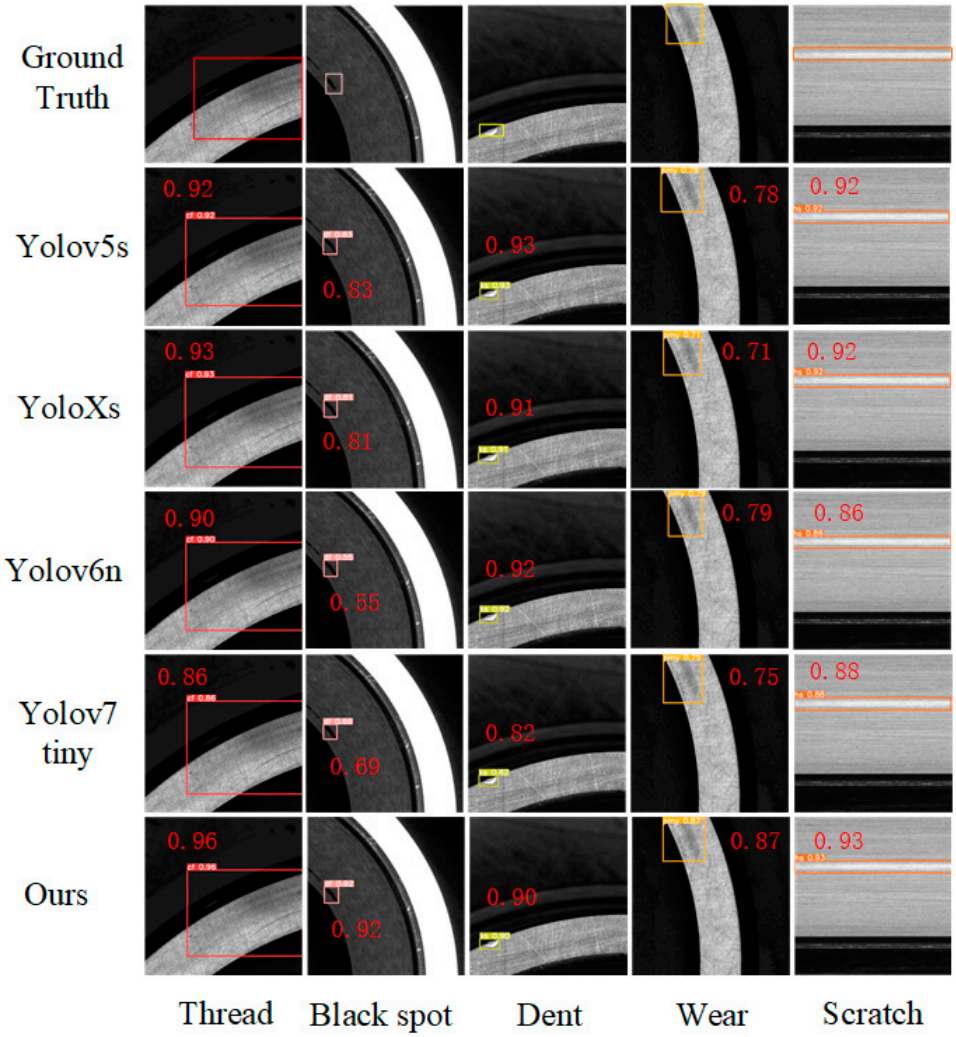


Figure 11. Test results of the different models on bearing collar defect dataset.

According to Figure 10, the loss curve of the ESD-YOLOv5 model rapidly converged within the first 30 epochs and achieved complete convergence after 100 epochs. The mAP curve also exhibited an increasing trend with the number of epochs. Compared to YOLOv5s, ESD-YOLOv5 showed faster convergence rates for all three losses, with the object loss exhibiting the most significant difference. The results also demonstrate that the ESD-YOLOv5 model achieves a higher mAP compared to YOLOv5s.

In this study, we compared our proposed ESD-YOLOv5 model with other single-stage object detection methods, including YOLOv5s, YOLOXs, YOLOv6n, YOLOv7tiny, and YOLOv8s, which have similar algorithm parameters and computational complexity. As shown in Table 7, both the YOLOv5s and YOLOv8s models achieved an mAP of 96.3%, which was the highest among all the original models. However, YOLOv5s has a lower parameter quantity and computational complexity compared to YOLOv8s. Our proposed ESD-YOLOv5 model achieved an mAP of 98.6%, which is a significant improvement of 2.3%. However, due to the increased parameters and computational complexity, the FPS of our proposed model slightly decreased. To ensure fairness, we conducted a comparative experiment with YOLOv5m. As shown in Table 6, ESD-YOLOv5 and YOLOv5m had similar FLOPs, but ESD-YOLOv5 achieved a higher mAP. Therefore, the proposed ESD-YOLOv5 model demonstrates better overall performance in terms of comprehensive evaluation metrics.

Five images were randomly selected for testing on each model, and the results are shown in Figure 11. It was observed that the different models have varying detection performances on the bearing collar defect dataset. Among all the original models, YOLOv5 and YOLOX had the best detection performance, while YOLOv6 and YOLOv7 had the poorest performance. All the original models had poor detection performance for black spots and wear, and the proposed ESD-YOLOv5 model improved the detection capability for these two defects.

5.5.2. Experimental Results of Hot-Pressed LGP and Fabric Datasets

To further verify the generality of the proposed ESD-YOLOv5 algorithm, we conducted a comparative experiment on the surface defect datasets of hot-pressed light guide plates and fabrics using the same experimental method as the bearing collar surface defect dataset mentioned above. The hot-pressed light guide plate dataset [44] is constructed from images of defective light guide plates, and the resolution of the sample images in the dataset is 416 × 416, with a total of 4111 images of defective light guide plates. The fabric dataset [45] is constructed from images of defective fabrics, with a resolution of 400 × 400 pixels for each sample image. The dataset comprises a total of 2764 images of defective fabrics. The detection results with networks such as YOLOv5s, YOLOXs, YOLOv6n, and YOLOv7tiny are shown in Table 8.

Table 8. Comparison of related methods on the hot-pressed LGP and fabric datasets.

Model	mAP@0.5	
	Hot-Pressed LGP	**Fabric**
YOLOv5s	97.8%	98.2%
YOLOXs	95.3%	98.0%
YOLOv6n	93.2%	96.8%
YOLOv7tiny	93.6%	97.4%
Ours	99.2%	99.1%

As shown in Table 8, our proposed model also achieved the highest detection accuracy on both the hot-pressed LGP and fabric datasets. These results demonstrate that the ESD-YOLOv5 model is effective in detecting surface defects in various datasets.

6. Discussion

In this study, ESD-YOLOv5 had the following advantages:

(1) By incorporating the ECCA module into the backbone network, the model's capability to extract features related to defects has been enhanced.
(2) Replacing the original neck of YOLOv5 with a slim neck has reduced the model's parameter quantity and computational load, while simultaneously improving its feature fusion capacity.
(3) The introduction of decoupled heads has significantly accelerated the convergence speed of the loss function and enhanced the detection accuracy.
(4) The experiment revealed that ESD-YOLOv5 achieved a 2.3% improvement in mAP compared to YOLOv5s, and it outperformed the current mainstream one-stage object detection algorithms.

Weaknesses and future research:

The bearing collar dataset used in this study was obtained from an industrial setting, and we only selected the five most common defect classes for detection, leaving many other defects undetectable.

Despite ESD-YOLOv5 achieving an mAP of 98.6%, instances of false negatives and false positives still exist, which are unacceptable in practical applications.

To better address these limitations, future research should focus on designing new algorithms for detecting uncommon defects. Additionally, for addressing false negatives and false positives, we should continue in-depth research on the dataset and improve the deficiencies of the model.

7. Conclusions

This study proposed a bearing collar surface defect detection method based on ESD-YOLOv5, which addresses the challenges of different shapes, sizes, and positions of bearing collar surface defects, as well as complex texture backgrounds. First, the ECCA module was introduced into the YOLOv5 backbone network to enhance the network's ability to locate object features. Second, the Slim-neck was used to replace the original neck, reducing the model's parameters and computational complexity without sacrificing accuracy. Third, the decoupled detection head of YOLOX was utilized to replace the original detection head, separating the classification and regression tasks. Last, extensive experiments were conducted on collected bearing collar defect images from industrial sites. The experimental results showed that the proposed algorithm achieved an mAP of 98.6% on the bearing collar defect dataset, with an overall improvement of 2.3%. In addition, we conducted experiments with our proposed ESD-YOLOv5 model on hot-pressed LGP and fabric datasets. The results demonstrated that our model also outperformed the current state-of-the-art one-stage object detection algorithms in terms of accuracy on two specific datasets. This further validates the superiority and versatility of our ESD-YOLOv5 model across different datasets and scenarios. Furthermore, the developed bearing collar defect detection system based on this method has been successfully applied in industrial production inspection.

Author Contributions: Conceptualization, J.L. (Jiale Li) and J.L. (Junfeng Li); methodology, J.L. (Jiale Li) and J.L. (Junfeng Li); software, J.L. (Jiale Li); validation, J.L. (Jiale Li) and J.L. (Junfeng Li); formal analysis, J.L. (Junfeng Li); investigation, J.L. (Jiale Li); resources, H.P.; data curation, J.L. (Jiale Li); writing—original draft preparation, J.L. (Jiale Li); writing—review and editing, J.L. (Junfeng Li); visualization, J.L. (Jiale Li); supervision, H.P.; project administration, H.P. and J.L. (Junfeng Li); funding acquisition, H.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key R&D Program of Zhejiang (No. 2023C01062) and Basic Public Welfare Research Program of Zhejiang Province (No. LGF22F030001, No. LGG19F03001).

Data Availability Statement: All data used in the experiments are from a private database. The datasets generated during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zheng, L.; Wang, X.; Wang, Q.; Wang, S.; Liu, X. A fabric defect detection method based on improved yolov5. In Proceedings of the 2021 7th International Conference on Computer and Communications (ICCC), Chengdu, China, 10–13 December 2021; IEEE: Washington, DC, USA.
2. Yao, J.; Li, J. AYOLOv3-Tiny: An improved convolutional neural network architecture for real-time defect detection of PAD light guide plates. *Comput. Ind.* **2022**, *136*, 103588. [CrossRef]
3. Li, W.; Zhang, H.; Wang, G.; Xiong, G.; Zhao, M.; Li, G.; Li, R. Deep learning based online metallic surface defect detection method for wire and arc additive manufacturing. *Robot. Comput.-Integr. Manuf.* **2023**, *80*, 102470. [CrossRef]
4. Gao, R.; Cao, J.; Cao, X.; Du, J.; Xue, H.; Liang, D. Wind Turbine Gearbox Gear Surface Defect Detection Based on Multiscale Feature Reconstruction. *Electronics* **2023**, *12*, 3039. [CrossRef]
5. Roy, A.M.; Bhaduri, J. DenseSPH-YOLOv5: An automated damage detection model based on DenseNet and Swin-Transformer prediction head-enabled YOLOv5 with attention mechanism. *Adv. Eng. Inform.* **2023**, *56*, 102007. [CrossRef]
6. Available online: https://github.com/ultralytics/yolov5 (accessed on 7 December 2022).
7. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:200410934.
8. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:220902976.
9. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:220702696.
10. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
11. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
12. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. *arXiv* **2022**, arXiv:220602424.
13. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:210708430.
14. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
15. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef] [PubMed]
17. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part I 14. Springer: New York, NY, USA.
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
19. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
20. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:180402767.
21. Available online: https://github.com/ultralytics/ultralytics (accessed on 20 July 2023).
22. Simonyan, K. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.556.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
24. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:170404861.
25. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
26. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
27. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
28. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
29. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.

30. Tang, Y.; Han, K.; Guo, J.; Xu, C.; Xu, C.; Wang, Y. GhostNetV2: Enhance Cheap Operation with Long-Range Attention. *arXiv* **2022**, arXiv:221112905.
31. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
32. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
34. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
35. Wang, X.; Yang, X.; Zhang, S.; Li, Y.; Feng, L.; Fang, S.; Lyu, C.; Chen, K.; Zhang, W. Consistent-Teacher: Towards Reducing Inconsistent Pseudo-Targets in Semi-Supervised Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 3240–3249.
36. Xu, B.; Chen, M.; Guan, W.; Hu, L. Efficient Teacher: Semi-Supervised Object Detection for YOLOv5. *arXiv* **2023**, arXiv:2302.07577.
37. Jiang, B.; Chen, S.; Wang, B.; Luo, B. MGLNN: Semi-supervised learning via multiple graph cooperative learning neural networks. *Neural Netw.* **2022**, *153*, 204–214. [CrossRef]
38. Zheng, Z.; Zhao, J.; Li, Y. Research on detecting bearing-cover defects based on improved YOLOv3. *IEEE Access* **2021**, *9*, 10304–10315. [CrossRef]
39. Lei, L.; Sun, S.; Zhang, Y.; Liu, H.; Xie, H. Segmented embedded rapid defect detection method for bearing surface defects. *Machines* **2021**, *9*, 40. [CrossRef]
40. Xu, J.; Zuo, Z.; Wu, D.; Li, B.; Li, X.; Kong, D. Bearing Defect Detection with Unsupervised Neural Networks. *Shock. Vib.* **2021**, *2021*, 9544809. [CrossRef]
41. Liu, B.; Yang, Y.; Wang, S.; Bai, Y.; Yang, Y.; Zhang, J. An automatic system for bearing surface tiny defect detection based on multi-angle illuminations. *Optik* **2020**, *208*, 164517. [CrossRef]
42. Fu, X.; Li, K.; Liu, J.; Li, K.; Zeng, Z.; Chen, C. A two-stage attention aware method for train bearing shed oil inspection based on convolutional neural networks. *Neurocomputing* **2020**, *380*, 212–224. [CrossRef]
43. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
44. Li, J.; Yang, Y. HM-YOLOv5: A fast and accurate network for defect detection of hot-pressed light guide plates. *Eng. Appl. Artif. Intell.* **2023**, *117*, 105529. [CrossRef]
45. Guo, Y.; Kang, X.; Li, J.; Yang, Y. Automatic Fabric Defect Detection Method Using AC-YOLOv5. *Electronics* **2023**, *12*, 2950. [CrossRef]