IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 September 2023, accepted 18 September 2023, date of publication 22 September 2023,
date of current version 28 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3318016

RESEARCH ARTICLE

A Lightweight Assembly Part Identification and Positioning Method From a Robotic Arm Perspective

LIGANG WU[1,2], LE CHEN[3], QIAN ZHOU[4], JIANHUA SHI[1], AND MINGMING WANG[1]
[1]College of Mechanical and Electrical Engineering, Shanxi Datong University, Datong 037003, China
[2]School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China
[3]College of Coal Engineering, Shanxi Datong University, Datong 037003, China
[4]College of Business, Shanxi Datong University, Datong 037003, China

Corresponding author: Ligang Wu (ligangwu@yeah.net)

This work was supported in part by the Shanxi Provincial Science and Technology Department Surface Project 202303021211330; in part by the Shanxi Province Higher Education Institutions Science and Technology Innovation Program Innovation Platform Project 2022P009; in part by the Shanxi Datong University 2022 Basic Research Fund Project 2022K1; in part by the Datong Key Research and Development Project 2020016; in part by the Shanxi Datong University Graduate Education Innovation Project 22CX35; in part by the 2023 China Society of Logistics, China Federation of Logistics and Purchasing Research Project 2023CSLKT3-237; and in part by the Shanxi Datong University General Program of Philosophy and Social Science under Grant 2022K06.

ABSTRACT To address the problems of low precision and poor real-time performance in the process of part identification and positioning of production line assembly robotic arm, Ghost-SE YOLOv5, an assembly part identification and positioning algorithm integrating lightweight network and attention mechanism is proposed. First, the redundancy of feature map convolution is utilized, which solves the problems of large number of model parameters and floating point operations by using Ghost convolution and Ghost Bottleneck modules. Second, the attention mechanism SE Module is introduced in the backbone network to increase the propensity of feature extraction. Last, the loss function is optimized to speed up the convergence of the model. The results shows that the number of parameters, float operation per second and train time of the proposed algorithm are reduced by 45.98%, 55.99% and 24.07%, respectively. And GPU use was reduced from 7.61G to 6.43G. Furthermore, during the test the precision reached 98.6%, and the recall rate realized 95.3%. The real-time detection performance achieved 97.59 FPS, with an improvement of 34.53%. It can be seen that Ghost-SE YOLOv5 algorithm has better practicality in the part identification and positioning of robotic arm for production line assembly.

INDEX TERMS Deep learning, lightweight, attention mechanisms, machine vision, real-time detection.

I. INTRODUCTION

It is proposed in ''Made in China 2025'' that ''high-end computerized numerical control (CNC) and robotics'' is listed as one of the ten key development areas, and assembly part identification and positioning is an important part of the CNC and flexible assembly robot (robot arm) vision tasks [1]. Traditional identification and analysis methods have low precision, slow efficiency and poor real-time performance. With the development of intelligent manufacturing and the breakthrough of key technologies of intelligent production lines in the new situation [2], it has become an indispensable part of developing Chinese manufacturing to solve the identification and positioning of assembly parts through deep learning and machine vision.

At present, machine vision technology and deep learning methods continue to develop and innovate, and performance continues to improve, in production line assembly robotic arm part identification and positioning, it is important to improve detection precision, reduce positioning errors, and achieve higher performance in real-time detection for enhancing the performance of intelligent manufacturing production lines [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

Machine vision technology and deep learning methods are not only applicable to the identification and positioning of assembly parts in the production process, but also have a wider application in the classification and recycling of disassembled parts. In Ref [4], [5], the authors achieved the identification and positioning of parts on the production line through the YOLOv3 target detection algorithm, and completed the auxiliary grasping of the intelligent assembly robot, but the detection performance in the paper is weak and does not reach the standard of real-time detection. In Ref [6] combined the YOLOv5 algorithm with a micro aircraft to achieve real-time detection and grasping of target objects on the Unmanned Aerial Vehicle robotic arm, solving the problem that embedded devices cannot achieve real-time detection. In Ref [7] that a surface defect detection method based on YOLOv7 is proposed based on full-dimensional dynamic convolution, but the accuracy is only 88.7% and the defect effect is poorly detected.

In Ref [8], the authors have effectively completed the defect detection of printed circuit boards by high-precision target detection. Similarly, in Ref [9], in order to solve the problem of recognition and positioning of mechanical parts in the assembly process, the authors optimization the Faster R-CNN algorithm, which improves the precision. Similarly, in Ref 10, the authors improved the YOLOv8 algorithm to achieve feature detection of aperture radar images in defense science and technology production lines, improving the detection accuracy to 98%.

In the recycling and utilization of dismantled parts in the production line, by combining machine vision technology and deep learning method, Borold et al. [11] realizes more efficient recognition and classification of automobile disassembled parts. Furthermore, In Ref [12], the authors analyze the strengths and weaknesses of YOLOv5 and YOLOv7. The authors analyze the strengths and weaknesses of YOLOv5 and YOLOv7, and prove that YOLOv7 is more advantageous in some small target detection, but has poor localization ability and some limitations, while YOLOv5 is based on prediction strategy and has better detection ability.

However, the existing deep learning recognition and localization algorithms have high hardware resource consumption, low real-time detection efficiency, computational redundancy and large model size, which are not suitable for embedded development and application of robotic arm vision system [13], [14].

To address the aforementioned problems, the Ghost-SE YOLOv5 proposed in this paper improves and optimizes on the basis of YOLOv5. While achieving a lightweight model, the detection precision, recall and real-time detection efficiency are improved, thus the main innovation and contributions of this paper are as follows:

(a) By adjusting the overall structure of the network, utilizing the redundancy characteristics of the convolutional feature map, using the lightweight Ghost Conv [15] and Ghost Bottleneck [16] significantly reduces the number of parameters, network layers, floating point computations (Flops) and training time, while improving GPU utilization and freeing up more hardware resource space.

(b) The channel attention mechanism Squeeze and Excitation Module [17] (SE Module) is introduced between feature extraction and feature fusion network, which not only improves the receptive field of feature extraction, but also can fuse more channel features. Thus the feature saliency and detection precision of the detected object can be improved.

(c) The use of Swish nonlinear activation function [18] to improve the neural network's expressiveness of the model. The loss function is also optimized to reduce the loss value of training and to speed up the convergence of the model [19].

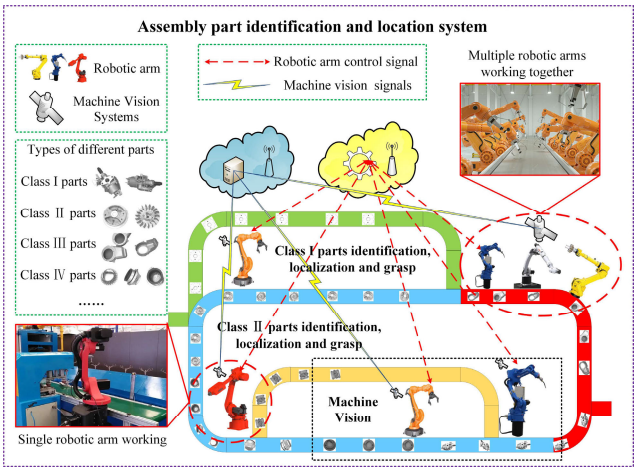


FIGURE 1. Assembly part identification and positioning system.

The remainder of this paper is organized as follows. Section II offers an assembly part identification and positioning system. In Section III, Ghost-SE YOLOv5 network structure and its constituent modules are presented. Sections IV introduce the object detection performance evaluation metrics, model training parameters and processes is provided, and analyze the results of comparison experiments. Finally, conclusions and future work are given in Section V.

II. IDENTIFICATION AND POSITIONING SYSTEM

During the assemble process of different products in the production line, the robot arm relies on computer vision technology to identify the assembled parts [20]. The precision and efficiency of assembly part identification directly affect the production efficiency of the assembly process, so a scientific and reasonable algorithm model, high detection precision and real-time detection performance are the prerequisites to meet the identification and positioning of assembly parts. It can be seen that the identification and positioning of assembly parts by deep learning methods and computer vision technology has important theoretical research significance and practical application value [21].

As shown in Figure 1, the assembly part identification and positioning system consists of two parts: assembly part identification and positioning based on machine vision technology and deep learning methods, robotic arm gripping and mounting. This research focuses on the identification and positioning of assembly parts. First, the machine vision system in the robot arm completes the identification and positioning of the assembled parts on the production line. Second, the robot arm is commanded to grip and install the assembly parts through a centralized control platform. Last, the robotic arms of different functions work together to complete the final assembly and production of the product.

During the production process, the working speed and overall smoothness of the robot arm are influenced by many factors, among them, mechanical speed and angle are influenced by their own properties, while the real-time detection speed and precision in the process of assembly part identification and positioning are influenced by the model performance and algorithm structure [22].

Therefore, this paper improves and optimizes the original YOLOv5 algorithm, and the Ghost-SE YOLOv5 algorithm is proposed that not only effectively compresses the training time and model volume, reduces the hardware resource consumption, and enables it to meet the requirements of embedded robotic arm vision system, but also has higher detection precision and high real-time detection speed compared with the original algorithm, which meets the requirements of identification and positioning of assembly parts in the industrial production process [23], [24].

III. GHOST-SE YOLOv5 ALGORITHM

The YOLOv5 target detection algorithm is directly trained end-to-end on the network, with excellent real-time performance and simple network structure, which is a more flexible algorithm in the current one stage algorithm, and has strong advantages in the field of multi-target detection and recognition [25]. Therefore, the proposed Ghost-SE YOLOv5 algorithm will be improved and optimized based on the YOLOv5 algorithm [26].

A. LIGHTWEIGHT CONVOLUTION AND MODULE

While the traditional convolution process generates many similar feature maps, lightweight convolution makes full use of the redundancy between feature maps, and achieves lightweight by reducing the redundant computation between feature maps [27].

As Figure 2 shows, Ghost Conv improves the efficiency of feature extraction, reduces the algorithm's demand and consumption of hardware resources with a cheap linear transformation operation. First, Ghost Conv generates m intrinsic feature maps by traditional convolution based on custom convolution kernels. Second, the m intrinsic feature maps are enhanced feature extraction performing the cheap linear transformation operations $\Phi(i < m)$, so that each intrinsic feature map generates $s-1$ new feature maps. Last, the m intrinsic feature maps generated by traditional

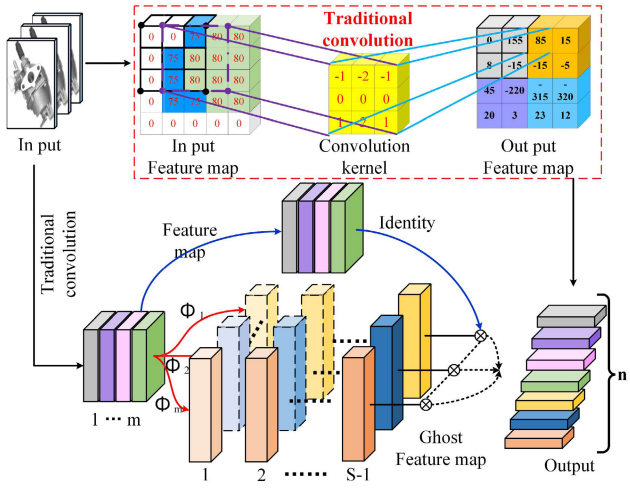


FIGURE 2. Ghost Conv and traditional convolution process.

convolution and the $s-1$ new feature maps generated by the cheap linear operation are stacked. At this point, the lightweight operation is completed, and a total of $m + m \times (s-1)$ feature maps are generated.

Suppose an input image and output image are $x \in R^{c\times h\times w}$ and $y \in R^{h'\times w'\times ms}$, respectively. Therefore, the calculation C_t of the traditional convolution can be calculated as

$$C_t = c \times k \times k \times ms \times h' \times w' \tag{1}$$

where, c is the number of input channels, ms is the number of output channels. $k \times k$ is the size of the custom convolution kernel, $h \times w$ and $h' \times w'$ are the height and width of the input and output images.

Ghost conv has the advantages of simple structure design, easy to operate, and can be used modularly. Therefore, the same sizes of convolution kernels, convolution strides and Padding are used in the Ghost Conv process as in the traditional convolution, which ensures that the size of the output feature map is the same as the traditional convolution. Thus, the number of computations required by Ghost Conv is

$$C_g = c \times k \times k \times \mathrm{m} \times h' \times w' + m \times k \times k \times (\mathrm{s}-1) \times h' \times w' \tag{2}$$

In formula (2), $c \gg s$ is usually satisfied. Therefore the theoretical speedup ratio r_s and model compression ratio r_c of Ghost Conv and traditional convolution are as follows:

$$\begin{aligned} r_s &= \frac{C_T}{C_G} \\ &= \frac{c\times k\times k\times ms\times h'\times w'}{c\times k\times k\times \mathrm{m}\times h'\times w' + m\times k\times k\times (\mathrm{s}-1)\times h'\times w'} \\ &= \frac{c\times s}{c+s-1} \approx s \end{aligned} \tag{3}$$

$$\begin{aligned} r_c &= \frac{c\times k\times k\times ms}{c\times k\times k\times m + \mathrm{m}\times k\times k\times (\mathrm{s}-1)} = \frac{c\times s}{c+s-1} \\ &\approx s \end{aligned} \tag{4}$$

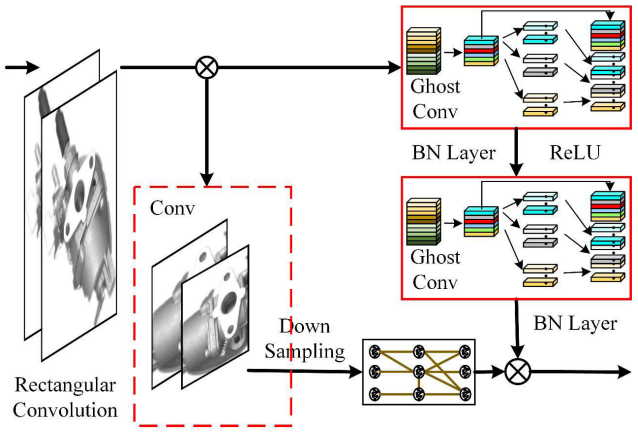


FIGURE 3. Ghost bottleneck module.

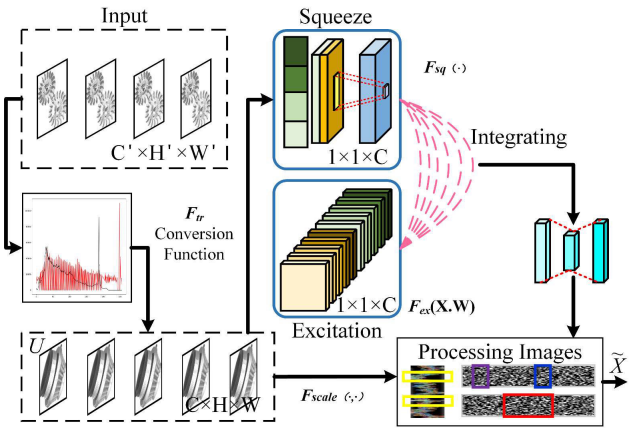


FIGURE 4. SE module structure.

In summary, Ghost Conv has more obvious advantages of the lightweight compared to traditional convolution. The calculation of Ghost Conv is about the $1/s$ of traditional convolution, which greatly reduces the computation of convolution process and significantly compresses the model volume and training time.

Ghost Conv can replace traditional convolution to reduce computation and compress model size. Thus, as shown in Figure 3, the lightweight module Ghost Bottleneck with Stride is 1 consists of Ghost Conv Module, Batch Normalization (BN) layer, down sampling and ReLU activation function.

The Ghost Bottleneck Module consists of two Ghost Modules stacked on top of each other [28]. The input image goes through the first Ghost Module to increase the number of channels, and through the BN layer and ReLU activation function by the second Ghost Module to reduce the number of channels, which ensures that the number of channels does unchanged, so it can be plug-and-play [29].

The proposed algorithm Ghost-SE YOLOv5 uses the Ghost Bottleneck module to replace the cross-stage partial network layers in the original algorithm, because fewer convolutional and BN layers are used, so the number of model network layers is less, and the amount of model computation and parameters is lower.

B. ATTENTION MECHANISM

In convolutional operations, the required information is obtained by extracting or fusing multi-scale spatial features. The introduction of the channel attention mechanism Squeeze and Excitation Module (SE Module) can improve the relationship between channels, adaptively adjust and correct the required feature weights in the feature extraction process, and the corrected channels will be more sensitive to the required feature information, thus improving the saliency of the detected object.

Remarks: X denotes the input tensor, $H' \times W'$ represents the height and width of the input image. C' indicates the number of input channels. U denotes the intermediate tensor, $\tilde{X}$ means the output tensor. $H \times W$ indicates the height and width of the input intermediate image and the output image, C represents the number of channels of the intermediate and output images, F_{tr} denotes conversion function.

Figure 4 shows the operation process of SE Module. First, the SE Module reduces the dimension of the input image, and performs global average pooling on the feature map of $H \times W \times C$, which achieves the purpose of compressing the information of each feature map. Subsequently, the global information is learned through feedforward networks and the corresponding weights of each feature map are obtained. Eventually, the obtained corresponding weights are multiplied with the original feature map to obtain the final feature information.

The operation of SE Module consists of two parts, Squeeze and Excitation. During the Squeeze process, the input image is adaptively average pool into the size of $1 \times 1 \times C$, the C dimensional vector is compressed to $1/r$ of the original, so the Squeeze operation is

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{5}$$

where, z_c denotes the output of the compression operation, F_{sq} represents the compression function. u_c is the c-th feature map, and (i, j) denotes the height and width of the c-th feature map.

The global information is obtained after compressing the model. Therefore, during the Excitation process, the nonlinear activation function ReLu is performed on the compressed C/r dimensional feature vector, and the C/r dimensional feature vector is raised to C dimension by the full connection operation. Thus, the Excitation process can be expressed as

$$s = F_{ex}(z, \mathrm{W}) = \sigma(\mathrm{g}(\mathrm{z}, \mathrm{W})) = \sigma(\mathrm{W}_2 \delta(\mathrm{W}_1 z)) \tag{6}$$

$$W_1 \in R^{\frac{C}{r} \times C}, \quad W_2 \in R^{C \times \frac{r}{C}} \tag{7}$$

where, F_{ex} denotes the excitation function, s denotes the output after excitation. z represents the output vector after the compression operation, which is also the input vector for the excitation operation. W denotes the mapping matrix, δ represents the activation function.

The channel attention mechanism SE Module Squeeze - Excitation process functions is as follows.

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c u_c \tag{8}$$

where, $\tilde{x}_c$ indicates the output of channel c after the image has passed through the SE Module, F_{scale} denotes the entire Squeeze-Excitation function, s_c denotes the weight of the c-th feature map.

In summary, the channel attention mechanism SE Module completes the adjustment of feature weights through feature Squeeze and Excitation, which is more conducive to obtaining the desired information.

C. OPTIMIZED LOSS FUNCTION

In the output prediction head of YOLOv5 target detection algorithm, the loss function consists of position error loss, classification loss and confidence loss. In the YOLOv5 algorithm, the category loss and confidence loss follow the BCE (binary cross-entropy) loss [30] function calculation method in YOLOv3 and YOLOv4, but the position error loss uses the generalized intersection over union (GIOU Loss) [31] function calculation method.

The calculation of GIOU Loss is shown in formula (9).

$$L_{GIOU} = 1 - GIOU = 1 - (IOU - \frac{|C - A \cup B|}{|C|}) \tag{9}$$

$$IOU = \frac{|A \cap B|}{|A \cup B|} \tag{10}$$

where, suppose A denotes the ground truth box of the image, B denotes the prediction box of the image, so $A \cap B$ represent the intersection of ground truth box and prediction box (grey shaded part), and $A \cup B$ represent the union of ground truth box and prediction box. C indicates the smallest external rectangular box of ground truth box and prediction box (red box).

The corresponding diagram analysis in the GIOU Loss is shown in Figure 5(a). GIOU Loss considers the case where ground truth box intersects with prediction box and has overlapping area. However, as shown in B_1, B_2, B_3 in Figure 5(a), when it appears that ground truth box contains prediction box and the size of prediction box is exactly the same, the A, B, C and difference sets in GIOU Loss are the same. In this case, it is impossible to distinguish the relative position relationship, which causes positioning errors and affects the detection precision.

For this reason, this paper uses the complete intersection over union (CIOU Loss) as shown in Figure 5(b) instead of the original loss function, taking into full consideration the overlap area, relative position relationship, centroid distance and the prediction box aspect ratio, which is calculated as

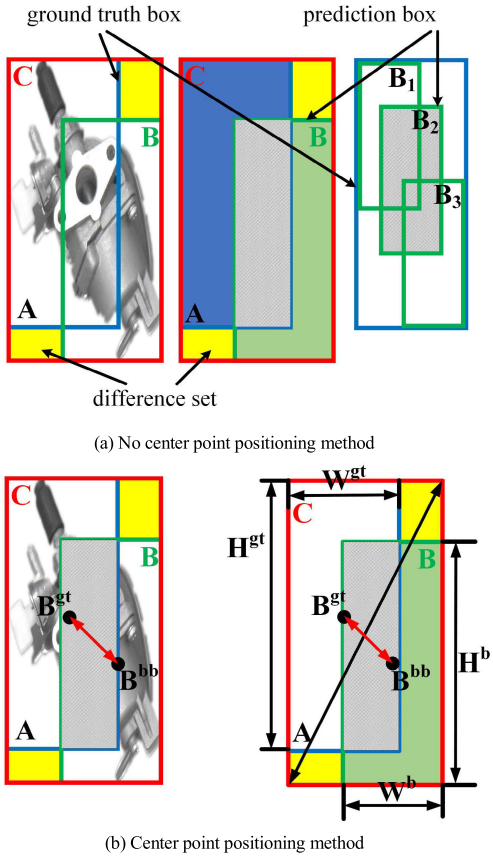


FIGURE 5. Schematic of loss function.

shown below.

$$L_{CIOU} = 1 - CIOU = 1 - (IOU - \frac{\rho^2(B^{gt}, B^{bb})}{C_P^2} - \frac{v^2}{(1 - IOU) + v}) \tag{11}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{W^{gt}}{H^{gt}} - \arctan\frac{W^b}{H^b}) \tag{12}$$

where, B^{gt} and B^{bb} denotes the center points of the rectangular boxes of ground truth box and prediction box, respectively. $\rho(B^{gt}, B^{bb})$ represent the Euclidean distance between two center points. C_P represent the diagonal distance from the minimum external rectangular box C of ground truth box and prediction box. W^{gt} and H^{gt} represent the width and height of the ground truth box, W^b and H^b represent the width and height of the prediction box.

It can be seen that CIOU Loss is more comprehensive than GIOU Loss, with more precision positioning and lower error, which is more in line with the requirements of production line assembly parts identification and positioning.

D. GHOST-SE YOLOv5 MODEL STRUCTURE

The structure of the improved lightweight neural network Ghost-SE YOLOv5 algorithm is shown in Figure 6. Ghost-SE YOLOv5 replaces the focus module with a 6 × 2 rectangular

convolution with Strict is 2, reducing the number of channels by 75%. At the same time, Ghost Conv and Ghost Bottleneck are used to replace the CBL and CSP (Cryptographic Service Provider) modules in the original algorithm, respectively, significantly reducing the number of parameters and Flops.

In the Ghost-SE YOLOv5 algorithm, the feature extraction backbone network completes down sampling through Ghost Conv and Ghost Bottleneck, and the Neck network completes feature fusion through Ghost Conv, C_3 residual module and up sampling.

With less convolution in the bottom network, it is able to contain more location and detail information, and has better feature resolution. However, in the high-level network, the resolution and detail perception are lower, but with stronger semantic information. Therefore, feature fusion networks can combine the underlying detailed information and high-level semantic information from the backbone network, each of which can be used to improve the model expressiveness performance.

Between the Backbone and Neck, adding the *SPP_F* module can make full use of the feature information extracted by the feature extraction backbone network, and introducing the channel attention mechanism SE Module can not only improve the feature extraction tendency of the model, but also the feature weight adjustment can be done adaptively.

The Ghost-SE YOLOv5 algorithm structure is shown in Figure 6. At the prediction output head of the network, three tensors of different sizes $(256, na \times (nc + 5))$, $(512, na \times (nc + 5))$ and $(1024, na \times (nc + 5))$ are generated, where 256, 512 and 1024 denote the number of input channels, $na \times (nc + 5)$ denotes the number of output channels. The number of anchors for each category and the number of categories of detected objects are denoted by na and nc, respectively. 5 denotes 4 localization parameters and 1 confidence parameter for each Anchor.

IV. MODEL TRAINING AND RESULT ANALYSIS

A. PERFORMANCE INDEXES

The performance evaluation metrics of the target detection model include precision (P), recall (R), $R-P$ curve, average precision (AP), and $mAP@0.5$, $mAP@0.5:0.95$, harmonized mean curve (F_1), and real-time detection performance.

1) PRECISION AND RECALL INDEX

P and R measure the precision and coverage of the model detection process, respectively. They are calculated as follows:

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

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

where, TP (True positive) and FN (False negative) indicates the part of the detected object that is correctly predicted and incorrectly predicted, respectively. The FP (False positive) indicates the part of the background object that is mistakenly detected as the target object.

A $R-P$ curve describing the average precision (AP) is formed by P and R, showing the variation trend of the algorithm model precision with recall during the training process. The $R-P$ curve can comprehensively evaluate the reliability of the model, and the larger area under the $R-P$ curve line, the higher the AP of the model.

The $mAP@0.5$ represents the average value of AP in N categories when the threshold IOU is taken as 0.5. Therefore, the $mAP@0.5$ is calculated as follows:

$$mAP@0.5 = \frac{1}{N}\sum_{i=1}^{N} AP_i(IOU_{th} = 0.5) \tag{15}$$

The $mAP@0.5:0.95$ defines the precision index under different values of the threshold IOU in the N categories, where j denotes the value taken during the change of the threshold from 0.5 to 0.95 with the stride is 0.05. It can be expressed as follows:

$$mAP@0.5:0.95 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j} AP(\mathrm{IOU}_{th} = j) \tag{16}$$

2) HARMONIC MEAN INDEX F_1

The F_1 comprehensively considers the P and R of the model, which is a harmonic mean calculation of the P and R. It is usually a constant between 0 and 1. When F_1 takes the maximum value, the robustness and comprehensive performance of the model is best at this confidence level.

$$F_1 = \frac{2 \times P \times R}{P + R + \varepsilon} \tag{17}$$

where ε is a negligible minimum value, usually taken as e^{-16}.

3) REALTIME DETECTION PERFORMANCE

Real-time detection performance reflects the detection capability of the trained model, real-time detection performance is measured by the number of image or video frames processed per second (FPS). The larger the FPS, the better the real-time detection performance and the higher the efficiency.

$$FPS = \frac{n}{t} \tag{18}$$

where, n denotes the number of frames processed of the image or video in time t. Generally, the real-time detection speed is requires to be greater than 30 FPS.

B. NETWORK MODEL TRAINING

The experiments in this paper are carried out using Python 3.8.5 environment and CUDA 11.3, under Intel Core i9-10900k@3.7GHz, NVidia GeForce RTX 3080 10G and DDR4 3600MHz dual memory hardware.

In this experiment, the image input is 2112×1419, the learning rate is 0.01, the cosine annealing hyper-parameter is 0.1, the weight decay coefficient is 0.0005 and the momentum parameter in gradient descent with momentum is 0.937.

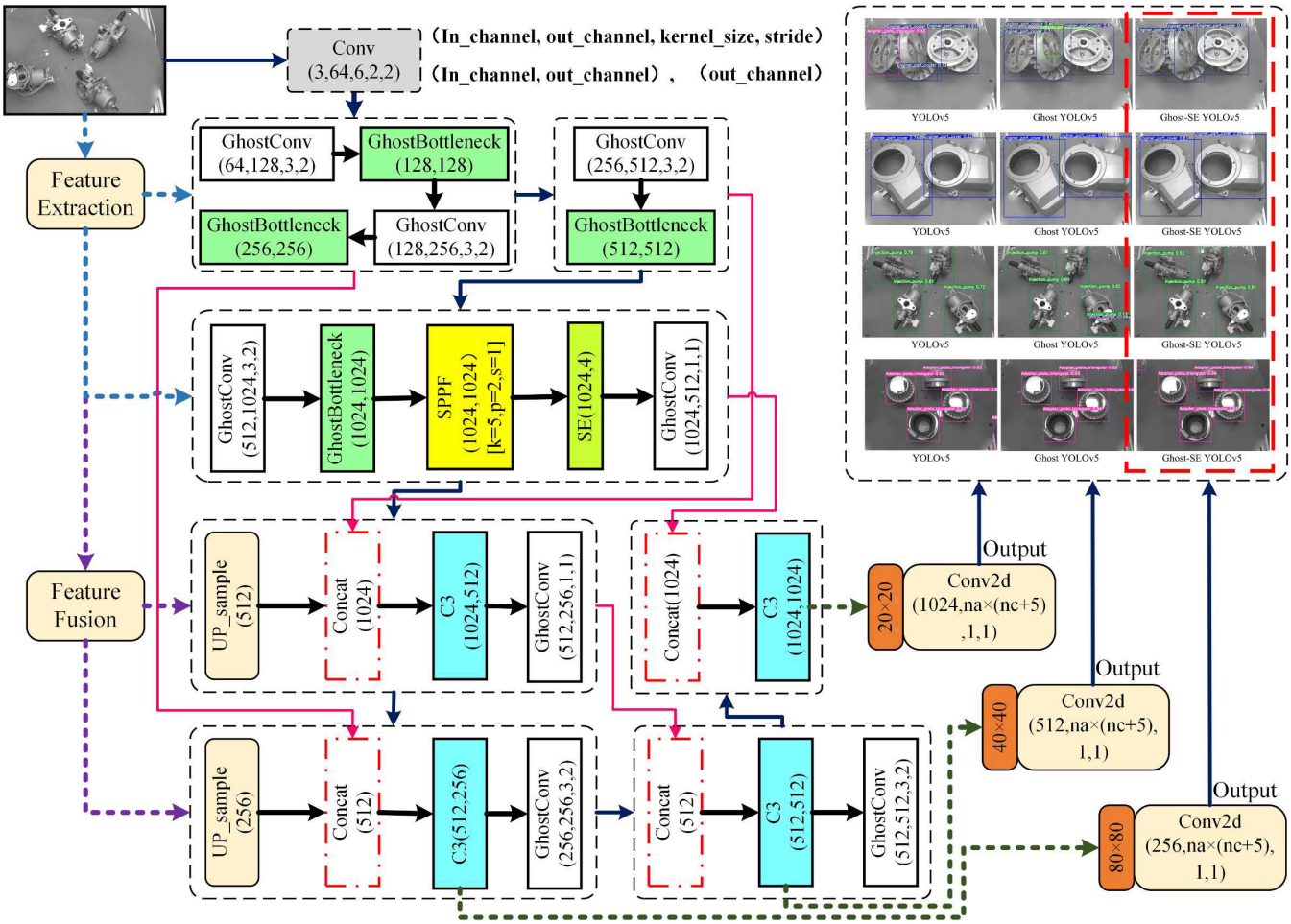


FIGURE 6. Ghost-SE YOLOv5 algorithm structure.

A total of 300 epochs and a batch size of 12 are used during training.

Based on the original YOLOv5 algorithm, the lightweight algorithm is named Ghost YOLOv5, which introduce Ghost Conv and Ghost Bottleneck. On this basis, the algorithm introduced the attention mechanism SE Module, and named Ghost-SE YOLOv5.

C. EXPERIMENTS

To verify the rationality and applicability of the proposed algorithm, experiments are conducted on Pascal VOC, T-LESS and MVTec ITODD datasets in this paper.

The Pascal VOC dataset is the standard dataset with high confidence for validating the proposed Ghost-SE YOLOv5 algorithm. The T-LESS and MVTec ITODD datasets contain 49 industrial assembly parts, which are masked from each other, and without obvious texture and color distinction, and thus can validate the utility of the model.

In this experiment, there are a total of 8832 images in the industrial dataset, including 5888 images in the training set, 2644 images in the validation set, and 300 images in the test set, and a total of 300 Epochs are trained iteratively. The training process of the improved and optimized Ghost-SE YOLOv5 algorithm is shown in Table 1.

1) IN PASCAL VOC STANDARD DATASET COMPARISON EXPERIMENT

Based on the YOLOv5 algorithm, Ghost-SE YOLOv5 achieves lightweight through Ghost Conv and Ghost Bottleneck. Subsequently, taking into full consideration the overlap area, relative position relationship, centroid distance and the prediction box aspect ratio of the ground truth box and prediction box, using a more reasonable CIOU Loss to complete the position fitting. Last, the SE Module is introduced to adjust the feature weights and speed up the convergence of the loss function. Therefore, the results of the ablation experiments on the Pascal VOC standard data set are shown in Table 2.

During the training process, the lower the value of the loss function, the higher the precision of the model, and vice versa. Figure 7 shows the loss function convergence curves and precision curves in Pascal VOC dataset. In the Figure 7(a), compared to the original algorithm, it can be observed that the proposed Ghost-SE YOLOv5 algorithm has

TABLE 1. Training process of Ghost-SE YOLOv5.

Ghost-SE YOLOv5 Algorithm
Determine: Parameters, Anchor, learning rate, loss.
InPut: Training dataset, Valid dataset, Label set.
Loading: Train models, Validate models.
Ensure: input, feature extraction, feature fusion, output.
N iterations of training. i-th iteration training($i \le N$):
Train Net:
a: Rectangular conv, Ghost Conv, SE AM, SPPF.
b: Predicted: classification c_i, confidence p_i.
c: error: Positioning, category, confidence.
d: aggregate CIOU losses.
Val Net:
a: Test effect of model ϖ_i.
b: Calculate P, R, mAP@0.5, mAP@0.5:0.95, FPS.
c: Adjust learning rate and update training strategy.
Save results of the i-th training: weight π_i, and model ϖ_i.
Update: Weight: $\pi_{i+1} \leftarrow \pi_i$,
Model: $\varpi_{i+1} \leftarrow \varpi_i$, temporary storage model ϖ'.
Save: Results, best model ϖ', **Output.**
End Train

lower loss function values at the early stage of training, which indicates that the Ghost-SE YOLOv5 has more adequate feature extraction and better information fusion capability, learning performance and advantages is more obvious at the pre-training stage. This is consistent with the precision curve in Figure 7(b), where the loss function values are low and the model precision is high in the early stage of the training process.

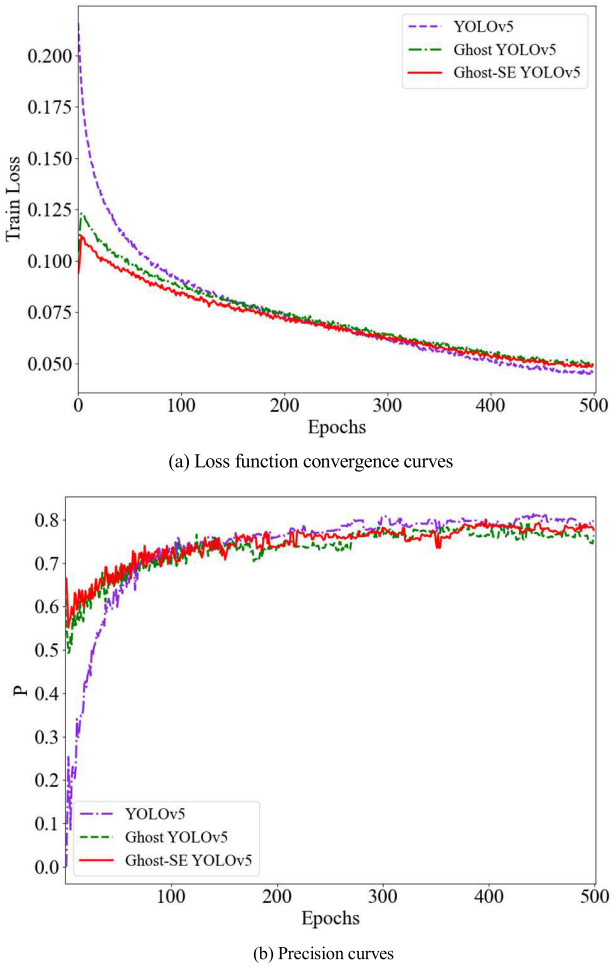
2) IN T-LESS AND MVTEC ITODD INDUSTRIAL ASSEMBLY PARTS DATASET COMPARISON EXPERIMENT

The identification and positioning of assembly parts in industrial production lines require accurate identification, correct positioning, and stable efficiency, meanwhile, to ensure production efficiency should have high real-time detection speed. So, to verify the reliability of the proposed algorithm needs to take into account both R and P.

The $R-P$ curve is an important performance metric, it reflects the negative correlation relationship between R and P. In T-LESS+MVTec ITODD assembly parts dataset, the $R-P$ curves of YOLOv5, Ghost YOLOv5 and Ghost-SE YOLOv5 algorithms during training are shown in Figure 8.

As Figure 8 shows, compared to the YOLOv5 and Ghost YOLOv5 algorithms, the Ghost-SE YOLOv5 algorithm has a larger area under the $R-P$ curve line, which shows that the proposed algorithm has a higher average precision (AP) and satisfies $AP_{\text{Ghost-SE YOLOv5}} > AP_{\text{Ghost YOLOv5}} > AP_{\text{YOLOv5}}$.

Figure 9 shows the average precision curves for different IOU thresholds, which demonstrates the average variation process of the precision of all the detected categories. It can be observed that in $mAP@0.5$ and $mAP@0.5:0.95$ curves, the Ghost-SE YOLOv5 algorithm has higher initial precision and faster rise in the pre-training period and eventually converges.



(a) Loss function convergence curves

(b) Precision curves

FIGURE 7. Loss function convergence curves and precision curves in Pascal VOC dataset.

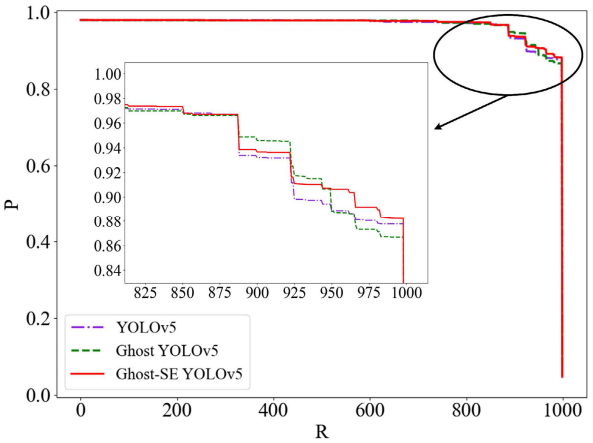
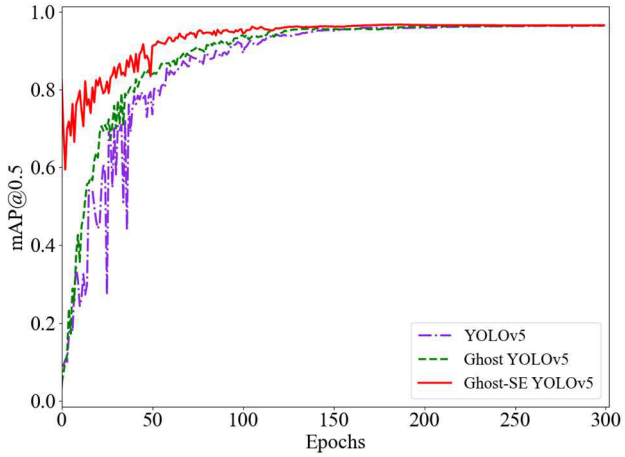


FIGURE 8. $R-P$ curves of different algorithms.

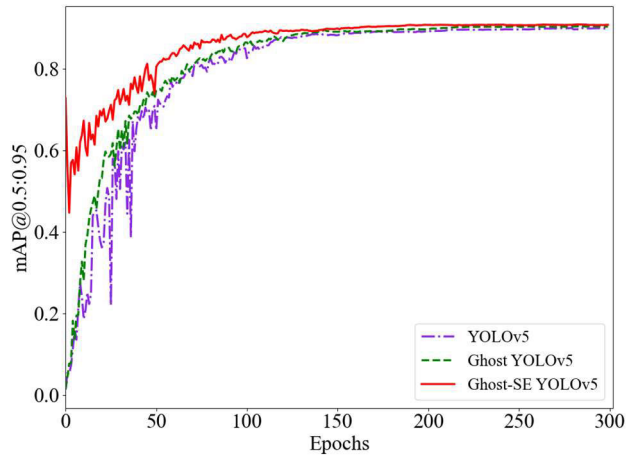
In the Figure 9(a), the final convergence value of $mAP@0.5$ curve of YOLOv5 algorithm is 0.9645, while the final convergence value of Ghost-SE YOLOv5 algorithm is 0.966, which is slightly higher than the original YOLOv5 algorithm. In the Figure 9(b), the convergence values of

TABLE 2. Ablation experiments.

YOLOv5	Lightweight Ghost Net	SE Module	GIOU Loss	Network Layers	P	FPS
√				499	0.736	38.01
√	√			351	0.734	46.98
√		√		505	0.739	42.17
√			√	499	0.737	40.53
√	√	√	√	357	0.744	45.83



(a) mAP@0.5 curves



(b) mAP@0.5:0.95 curves

FIGURE 9. mAP@0.5 and mAP@0.5:0.95 curves.

YOLOv5 and Ghost-SE YOLOv5 in the $mAP@0.5:0.95$ curve are 0.8987 and 0.9077, respectively, and the proposed Ghost-SE YOLOv5 algorithm is about 1% higher than the original algorithm.

In summary, it can be seen from AP, $R-P$ curves, $mAP@0.5$ curves and $mAP@0.5:0.95$ curves that the Ghost-SE YOLOv5 algorithm is higher than the original YOLOv5 algorithm in terms of precision and recall, which proves that the proposed algorithm has better practicality for assembly part recognition and localization based on machine vision and deep learning.

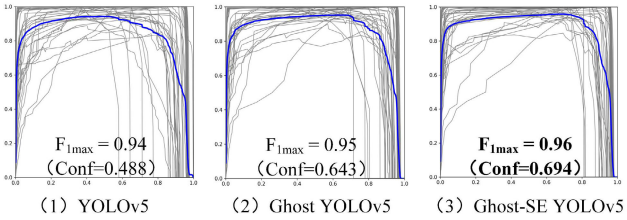


FIGURE 10. F_1 curves of different algorithms.

The harmonic mean performance curve shows the trend of harmonic calculation of model precision and recall during training with different confidence thresholds. During the training process, the F_1 harmonic mean curves of YOLOv5, Ghost YOLOv5 and Ghost-SE YOLOv5 algorithms are shown in Figure 10.

As Figure 10 shows, during the change of the confidence threshold from 0 to 1, YOLOv5 algorithm obtains the maximum value $F_1 = 0.94$ at confidence threshold $Conf = 0.488$, and Ghost-SE YOLOv5 achieves the maximum value $F_1 = 0.96$ at the confidence level $Conf = 0.694$. It can be seen that with a confidence level is 0.694, the trade-off calculation of P and R of Ghost-SE YOLOv5 algorithm is more reasonable, and the comprehensive performance of the model is more desirable.

D. ANALYSIS OF RESULTS

The confusion matrix represents the one-to-one quantitative statistical correspondence between the model true labels and the predicted category labels. The horizontal coordinate of the confusion matrix is the true label of the detected object, the vertical coordinate is the predicted class label, so the diagonal line of the confusion matrix indicates the relationship of the number of correct predictions during the training and validation of the model, where the darker color indicates the higher precision of the prediction [32].

As shown in Figure 11, in this paper a total of 49 industrial parts and 1 background can be detected, for a total of 50 categories to be detected, and all the backgrounds are considered as 1 category. Therefore, the confusion matrix is 50×50 and is one-to-one correspondence. Each of these

TABLE 3. Parameters of each models.

Model	GPU/G	Training times/s	mAP@0.5	mAP@0.5:0.95	F_1	Test time/300 images	*FPS*
YOLOv5	7.61	58924	0.965	0.899	0.94	4.695s	63.90
Ghost YOLOv5	**6.14**	**45803**	**0.965**	**0.903**	**0.95**	**2.897s**	**103.56**
Ghost-SE YOLOv5(**Our**)	**6.43**	**44741**	**0.966**	**0.908**	**0.96**	**3.074s**	**97.59**

TABLE 4. Comparison of other algorithms.

Algorithms	YOLOv5	Ghost-SE YOLOv5	YOLOv7	YOLOv8
P	96.7%	98.6%	95.4%	98.7%
R	94.9%	95.3%	92.7%	93.6%
mAP@0.5	96.5%	96.6%	97.0%	96.8%
mAP@0.5:0.95	89.9%	90.8%	87.1%	88.3%
F_1	0.94	0.96	0.95	0.92
Real time performance	63.90	97.59	80.57	89.53

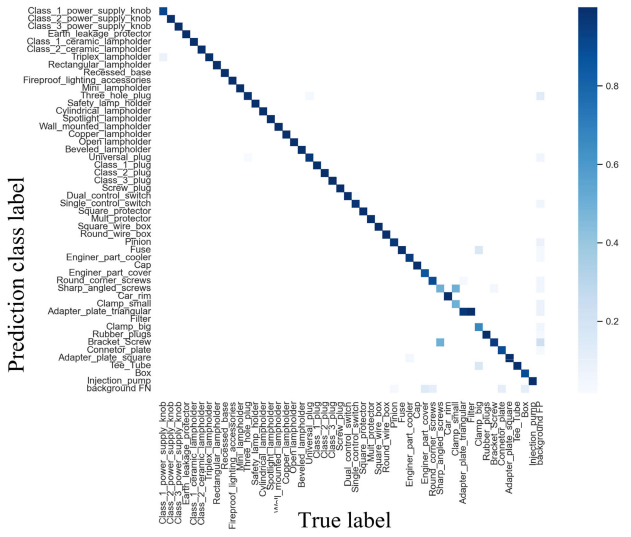


FIGURE 11. Confusion matrix.

categories has the potential to be detected as one of the other 49 categories, including the background (false detection).

As shown in the figure, taking the first column as an example, it can only be considered as accurately detected if it is detected as a category tagged by the label, i.e., the same as the category in the first row, otherwise it is a false detection.

As can be seen from the figure, the assembly part recognition based on machine vision and deep learning has high detection precision, there are very few false detections, but still can meet the needs of actual production.

As shown in Figure 12, during the training process with the T-LESS+MVTec ITODD dataset, due to the lightweight achieved by the Ghost-SE YOLOv5 algorithm, so the number of parameters, Flops, and model volume are reduced significantly compared to the YOLOv5 algorithm. Among

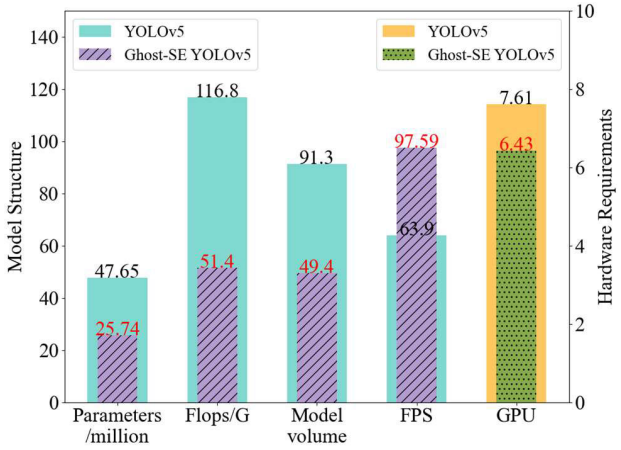


FIGURE 12. Training process comparison of different parameters.

them, the number of parameters and Flops are reduced by about 22 million and 65.4 G respectively, and the model volume is compressed by about 41.9 M.

However, the real-time detection performance of the improved Ghost-SE YOLOv5 improved significantly, from 63.9 FPS to 97.59 FPS, an improvement of about 34.19 FPS. In addition, the GPU occupancy is significantly lower, from 7.61G to 6.43G, a reduction of about 1.18 G.

Table 3 shows the parameters and test results of different algorithm models. It is clear from the table that Ghost-SE YOLOv5 not only reduces the GPU occupancy, but also the training time by about 24.07%. Furthermore, the real-time detection speed of 97.59 *FPS* is achieved, which is a 34.53% improvement. Which is much greater than the 63.9 *FPS* in the original algorithm and the 30 *FPS* required in real-time detection. It can be seen that Ghost algorithm has a strong advantage in detection precision and speed, and can meet the requirements of assembly part identification.

The analysis of the experimental results shows that Ghost YOLOv5 achieves a real-time detection performance of 103.56 *FPS*, because it introduces a light weight to optimize the convolution process and model structure. However, after the introduction of SE Module, the amount of computation and the number of network layers are increased, so the real-time detection performance is reduced to 97.59 *FPS*, but the significance of the detected target increases and the detection precision improves, due to the tendency of the channel attention mechanism in the feature extraction process.

In addition, experiments were conducted in three different algorithms, YOLOv5, YOLOv7 and YOLOv8, for assembly

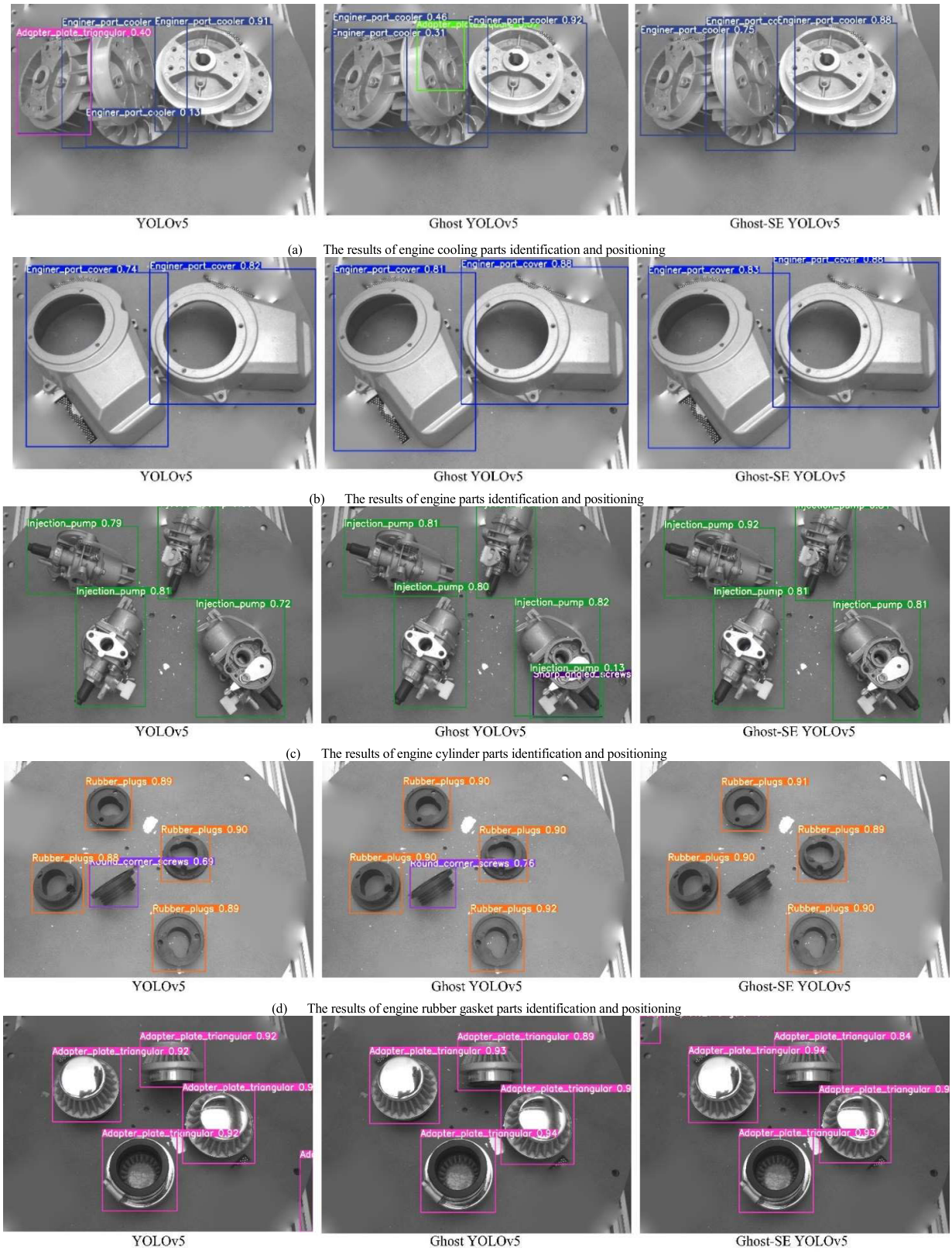


FIGURE 13. Test results of different assembly parts.

part identification and localization, and the results are shown in Table 4.

As shown in the table, compared to Ghost-SE YOLOv5, the YOLOv7 algorithm has lower real-time performance

and recall due to the addition of non-maximal suppression. Whereas YOLOv8 applies the unanchored network detection header and the new loss function, so the precision and mAP@0.5 are higher, but the real-time performance is worse.

Finally, in order to verify the feasibility of the algorithm proposed in this paper, different assembly parts were selected for testing and validation under multiple scenarios, and the comparison of the test results is shown in Figure 13.

In conclusion, the detection results show that the Ghost-SE YOLOv5 algorithm locates more accurately, the recall rate and label detection confidence are higher in the process of assembly part identification and localization. In the case of highly dense and highly overlapping, the missed detection rate is lower and the detection effect is significantly better than the original algorithm.

V. CONCLUSION

In this paper, in order to solve the problem of slow identification and localization of assembly parts under robot arm vision on production lines, we propose Ghost-SE YOLOv5, a recognition method that incorporates Ghost lightweight neural networks and SE attention mechanisms. In addition, the loss function is also optimized to effectively improve the precision and real-time performance of the model.

Compared with the original algorithm, the improved Ghost-SE YOLOv5 loss function converges faster, the number of network layers is reduced by about 28.46%, the precision is improved by about 0.8%, and the real-time performance is improved from 63.90 FPS to 97.59 FPS. The experimental results show that the proposed algorithm can meet the identification and positioning of production line assembly parts.

In future research work, we will try to combine deep learning with reinforcement learning, which will enable the model to have the ability of self-correction during the learning process and thus further improve the model performance.

REFERENCES

[1] X. Cui, L. Li, Y. Wang, L. Zhou, and F. Chang, "'Made in China 2025' background of high-end technical skills talents training innovation model research," in *Proc. 8th Int. Conf. Instrum. Meas., Comput., Commun. Control (IMCCC)*, Jul. 2018, pp. 666–670.

[2] A. A. Malik, M. V. Andersen, and A. Bilberh, "Advances in machine vision for flexible feeding of assembly parts," *Procedia Manuf.*, vol. 38, pp. 1228–1235, 2019.

[3] M. Javaid, A. Haleem, R. P. Singh, S. Rab, and R. Suman, "Exploring impact and features of machine vision for progressive industry 4.0 culture," *Sensors Int.*, vol. 3, 2022, Art. no. 100132.

[4] Y. W. Yu, X. Peng, L. Q. Du, and T. H. Chen, "Real-time detection of parts by assembly robot based on deep learning framework," *Act Armamentaria*, vol. 41, no. 10, pp. 2122–2130, 2020.

[5] J. Zhang, F. L. Liu, and R. W. Wang, "Research on industrial parts recognition algorithm based on YOLOv3 in intelligent assembly," *J. Optoelectronics Laser*, vol. 31, no. 10, pp. 1054–1061, 2020.

[6] R. Zhang, Y. Y. Wang, Y. Q. Duan, and B. Chen, "Real-time object detection and location algorithm for aerial manipulator," *J. Nanjing Univ. Aeronaut. Astronaut.*, vol. 54, no. 1, pp. 27–33, 2022.

[7] X. Hong, F. Wang, and J. Ma, "Improved YOLOv7 model for insulator surface defect detection," in *Proc. IEEE 5th Adv. Inf. Manage., Communicates, Electron. Autom. Control Conf. (IMCEC)*, Dec. 2022, pp. 1667–1672, doi: 10.1109/IMCEC55388.2022.10019873.

[8] L. Wu, L. Zhang, and Q. Zhou, "Printed circuit board quality detection method integrating lightweight network and dual attention mechanism," *IEEE Access*, vol. 10, pp. 87617–87629, 2022.

[9] G. Fei, W. Y Jin, and M. Wang, "Image recognition of mechanical parts based on the improved Faster R-CNN algorithm," *J. Mach. Des.*, vol. 36, no. 9, pp. 113–116, 2019.

[10] Y. Zhou, W. Zhu, Y. He, and Y. Li, "YOLOv8-based spatial target part recognition," in *Proc. IEEE 3rd Int. Conf. Inf. Technol., Big Data Artif. Intell. (ICIBA)*, Chongqing, China, May 2023, pp. 1684–1687.

[11] Z. Liu, X. Gao, Y. Wan, J. Wang, and H. Lyu, "An improved YOLOv5 method for small object detection in UAV capture scenes," *IEEE Access*, vol. 11, pp. 14365–14374, 2023.

[12] T. Reddy Konala, A. Nammi, and D. Sree Tella, "Analysis of live video object detection using YOLOv5 and YOLOv7," in *Proc. 4th Int. Conf. Emerg. Technol. (INCET)*, May 2023, pp. 1–6.

[13] J. Wu, J. Zhu, X. Tong, T. Zhu, T. Li, and C. Wang, "Dynamic activation and enhanced image contour features for object detection," *Connection Sci.*, vol. 35, no. 1, Dec. 2023, Art. no. 2155614.

[14] C. Jiang and J. Wan, "A thing-edge-cloud collaborative computing decision-making method for personalized customization production," *IEEE Access*, vol. 9, pp. 10962–10973, 2021.

[15] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1577–1586.

[16] Y. Tang, K. Han, J. Guo, C. Xu, and Y. Wang, "GhostNetv2: Enhance cheap operation with long-range attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 9969–9982.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2017.

[18] P. Ramachandran et al., "Searching for activation functions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[20] W. Landi and Z. Wei, "Basic modes and reform strategies of university-industry collaboration for made in China 2025," in *Proc. 7th World Eng. Educ. Forum (WEEF)*, Nov. 2017, pp. 140–144, doi: 10.1109/WEEF.2017.8467076.

[21] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[22] Q. Xiao, Q. Li, and L. Zhao, "Lightweight sea cucumber recognition network using improved YOLOv5," *IEEE Access*, vol. 11, pp. 44787–44797, 2023, doi: 10.1109/ACCESS.2023.3272558.

[23] X. G. Yang, F. Gao, and R. T. Lu, "Lightweight aerial object detection method based on improved YOLOv5," *Inf. Control*, vol. 51, no. 3, pp. 361–368, 2022.

[24] M. Zhang, H. Shi, Y. Zhang, Y. Yu, and M. Zhou, "Deep learning-based damage detection of mining conveyor belt," *Measurement*, vol. 175, Apr. 2021, Art. no. 109130.

[25] Z. Jiao, G. Jia, and Y. Cai, "A new approach to oil spill detection that combines deep learning with unmanned aerial vehicles," *Comput. Ind. Eng.*, vol. 135, pp. 1300–1311, Sep. 2019.

[26] L. Zhang, L. Wu, and Y. Liu, "*HemerocalliscitrinaBaroni* maturity detection method integrating lightweight neural network and dual attention mechanism," *Electronics*, vol. 11, no. 17, p. 2743, Aug. 2022, doi: 10.3390/electronics11172743.

[27] L. A. Estrada-Jimenez, T. Pulikottil, S. Nikghadam-Hojjati, and J. Barata, "Self-organization in smart manufacturing—Background, systematic review, challenges and outlook," *IEEE Access*, vol. 11, pp. 10107–10136, 2023.

[28] T. A. Zhou, *Lightweight Improvement of YOLOv5 for Insulator Fault Detection*. Bristol, U.K.: IOP Publishing, 2023.

[29] L. Yang, H. Cai, X. Luo, J. Wu, R. Tang, Y. Chen, and W. Li, "A lightweight neural network for lung nodule detection based on improved ghost module," *Quant. Imag. Med. Surg.*, vol. 13, no. 7, pp. 4205–4221, Jul. 2023.

[30] C.-Z. Wang, X. Tong, J.-H. Zhu, and R. Gao, "Ghost-YOLOX: A lightweight and efficient implementation of object detection model," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 4552–4558.
[31] K. M. He, X. Y Zhang, S. Q. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2015, pp. 346–361.
[32] X. Hu, J. Wan, T. Wang, and Y. Zhang, "An IoT-based cyber-physical framework for turbine assembly systems," *IEEE Access*, vol. 8, pp. 59732–59740, 2020, doi: 10.1109/ACCESS.2020.2983123.

LIGANG WU was born in Datong, Shanxi, China, in 1986. He received the M.Sc. and Ph.D. degrees in control theory and control engineering from Dalian Maritime University, Dalian, China, in 2012 and 2016, respectively. He is currently an Associate Professor with the College of Mechanical and Electrical Engineering, Shanxi Datong University. His current research interests include machine learning algorithms, intelligent control, intelligent manufacturing, and other artificial intelligence fields.

LE CHEN was born in Zhongxian, Chongqing, China, in 1998. She received the B.E. degree in communication engineering from Chongqing Three Gorges University, Chongqing, in 2021. She is currently pursuing the Graduate degree in source and environment with Datong University, Shanxi, China. Her research interests include machine learning algorithms, object detection, recognition, and localization.

QIAN ZHOU was born in Nanchong, Sichuan, China, in 1988. She received the M.Sc. degree in business management from the Southwestern University of Finance and Economics, Chengdu, China, in 2014. Her current research interest includes data processing.

JIANHUA SHI was born in Datong, China, in June 1978. He received the B.S. degree in automation, the M.Sc. degree in control theory and control engineering, and the Ph.D. degree in weapons systems and applications engineering from the North University of China, Taiyuan, China, in 2000, 2006, and 2014, respectively. He is currently a Professor with the School of Shanxi Datong University. His current research interests include deep learning, smart manufacturing and smart factories, and other artificial intelligence areas.

MINGMING WANG was born in Shuozhou, China, in May 1980. He received the M.Sc. degrees in mechanical design and theory from the Taiyuan University of Technology, Taiyuan, China, in 2010. He is currently a Lecturer with the College of Coal Engineering, Shanxi Datong University. His current research interests include machine learning intelligent manufacturing and mechatronics.