IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 November 2022, accepted 4 December 2022, date of publication 12 December 2022,
date of current version 20 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3228687

RESEARCH ARTICLE

Surface Defect Detection of Industrial Parts Based on YOLOv5

HAI FENG LE[1], LU JIA ZHANG[1], AND YAN XIA LIU[2]
[1]Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China
[2]College of Urban Rail Transit and Logistics, Beijing Union University, Beijing 100101, China

Corresponding author: Yan Xia Liu (yanxia.liu@163.com)

This work was supported in part by the Science and Technology Program of Beijing Municipal Education Commission under Grant KM201911417007, and in part by the key Program of Beijing Union University for Educational Reform under Grant JY2021Z002.

ABSTRACT Industrial product quality inspection, a crucial procedure in industrial production, is crucial in assuring product yield. Product safety and quality inspections on industrial assembly lines are predominantly manual, and there is currently a dearth of safe and dependable inspection techniques. An improved surface defect detection approach based on YOLOv5 is proposed for the problem of surface flaws in industrial components in order to improve the quality detection effect of industrial production parts. To improve the effect of dense object detection, the image features are extracted by the convolutional network and enhanced by coordinate attention. BiFPN is utilized to fuse multi-scale features in order to lower the rate of missed detection and false detection for small target samples. The detectors from the Transformer structure are added to the complex problem of fine-grained detection to improve the predictability of challenging occurrences. According to the experimental findings, on the dataset for industrial parts defects, the proposed network increases the recall of the original algorithm in abnormal classes by 5.3%, reaching 91.6%. Its inference speed can approach 95FPS, indicating an improved real-time detection performance.

INDEX TERMS Defect detection, YOLOv5, transformer, deep learning, fine-grained detection.

I. INTRODUCTION

The industrial product safety and quality inspection of industrial assembly lines are mainly based on manual review. Industrial production still relies on the naked eye to detect and analyze defective products on assembly lines. Considering the safety of human inspectors and the low efficiency of manual inspection of products on assembly lines, the most widely used method is to improve the recall rate through manual sampling after production [1]. This manual detection method is inefficient and has potential safety risks, which limits the update and development of the industrial chain in the long run. Hence, it is imminent to realize online detection of product defects on industrial assembly lines.

With the development of deep learning in recent years, computer vision has become more and more powerful and functional in addressing tasks such as image classification and object detection, characterized with increased recognition accuracy surpassing human eyes and increasingly faster recognition speed. Diverse algorithms and advanced computing equipment provide stable conditions for accurate industrial defect detection [2]. Robots and mechanical intelligence technology play an important role in industrial manufacturing, and defect detection methods used for related parts also have research significance [3], [4], [5]. Since some industrial defects can be regarded as the abnormal appearance of industrial products, it is suitable to adopt image methods for detecting abnormality [6], [7]. In particular, image anomaly detection mainly focuses on whether the input image is an anomaly instance, usually at the image level. However, industrial defect detection is more concerned with detection tasks at the local level. At the local image and pixel level, the difference between anomalies and standard patterns is more subtle, which leads to significantly increased difficulty in actual detection. Therefore, employing image-level anomaly detection methods to meet the requirements of industrial defect detection is challenging. Yue et al. [8] propose to use a deep learning method for local defect detection

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

of industrial products. Yang et al. [9] locate various defect positions through the object detection method and distinguish defect categories using an improved classification network. Zhao et al. [10] extract defect information by virtue of the instance segmentation method. The prediction results are output through the subsequent network, and the training data is enriched with weakly supervised learning. Still, real-time detection cannot be guaranteed due to slow recognition speed. The industrial field requires real-time detection performance, and even small embedded devices can satisfy the real-time detection requirements. This application scenario requires a lightweight, high detection frame rate model.

To solve the problems of low detection accuracy and inability to real-time detection in traditional methods [11], [12], a mechanical product defect detection system is proposed in this paper for industrial assembly lines based on the object detection method. Parts with defects such as deform and contamination are marked when their appearance is detected to be defective, which provides convenience for subsequent early warning and rejection of defective products. Different from other object detection methods in defect detection of industrial parts, this paper detects different abnormal states of the same category at the instance level, which belongs to fine-grained detection, characterized by the distinction between different abnormal categories. Due to smaller difference of the samples in different states, it is more difficult to identify correct samples. Using a lightweight model with high real-time performance, the proposed method can provide a computer vision-based solution for current industrial production through embedded transplant deployment, so as to enhance the quality of products under industrial assembly lines. Based on ensuring real-time detection performance, given the difficulty of defect classification with the existing YOLOv5s method for small samples with low recall rates and slight sample differences, the model is improved in this paper to make it more suitable for tiny target and difficult sample detection to obtain satisfactory results. The main contributions are:

1) The feature extraction module can be given coordinate attention to significantly improve the detection performance of the model with minimal computational overhead.
2) By using a bidirectional multi-scale fusion module, it is possible to optimize the model hierarchy, fuse additional layers of features without increasing extra calculation. It can also enhance the feature fusion ability of the network, and raise the recall rate for small target samples.
3) Aiming at the issue of missing detection of fine-grained samples in the dataset, a detector with a Transformer structure is proposed to enhance the feature extraction capability of the model and effectively increase the recognition accuracy of difficult and difficult target samples.

The remainder of this paper is organized as follows. In Section 2, we introduce related works on object detection in recent years. Section 3 presents the details of the proposed method. The implementation of the proposed method and comparison with previous methods is presents in Section 4. Section 5 summarizes the conclusions of the work in this study and suggests the future search direction.

II. RELATED WORK

Object detection includes two parts, classification and location, and its application fields are broad, including face detection, pedestrian detection, vehicle detection, etc [13]. Traditional object detection algorithms adopt sliding windows to detect objects without any pertinence, which is inefficient and inaccurate. The manually selected features are less robust to irregular objects with different shapes [14]. With the advancement of deep learning technology, image feature extraction by the convolutional neural network has become a common approach [15], [16], [17]. Meanwhile, object detection, as one of the hot spots in the field of machine vision research, has stimulated the appearance of numerous excellent algorithms in object detection [18], [19], [20]. The emergence of abundant networks has played a critical role in promoting the development of deep learning. For example, ResNet [21] proposed the concept of residual blocks, which significantly intensified the depth of networks. Also, new feature extraction methods are provided in terms of image detection with the help of the attention mechanism [22]. Methods, such as the assemblable attention module proposed by SENet [23], bring accuracy improvement to the convolutional network; DETR [24] uses the classical convolution structure to encode the image features after extraction and completes both classification and positioning through the transformer structure. For detection, an innovative Hungarian loss function is used to match the decoded target class in general detection networks, rather than the initial anchor design. Similar to natural language processing, VIT [25] encodes the segmented and serialized images to input them into the transformer, and directly obtains coordinate positions and category of targets through encoding and decoding. Swin-transformer [26] is improved based on VIT, which can solve the problem of the enormous computational cost of the VIT method through hierarchical feature mapping and window attention transformation.

So far, two main branches of object detection methods are mentioned on the basis of deep learning: the two-stage object detection model based on the region generation network and the one-stage object detection model that directly performs position regression [27]. YOLOv5 is an efficient and stable one-stage object detection method with greatly enhanced speed and accuracy, and can quickly adapt to new tasks after transfer learning. The input of the YOLOv5 is an RGB image with a size of 640*640. Its overall network design is divided into a backbone network based on the CSPNet [28] neural network, a multi-scale feature fusion module based on the FPN [29]+PAN [30] structure and the detector for output classification and bounding box regression.

The backbone of YOLOv5 includes Focus, BottleneckCSP and SPP. The first two components mainly undertake image

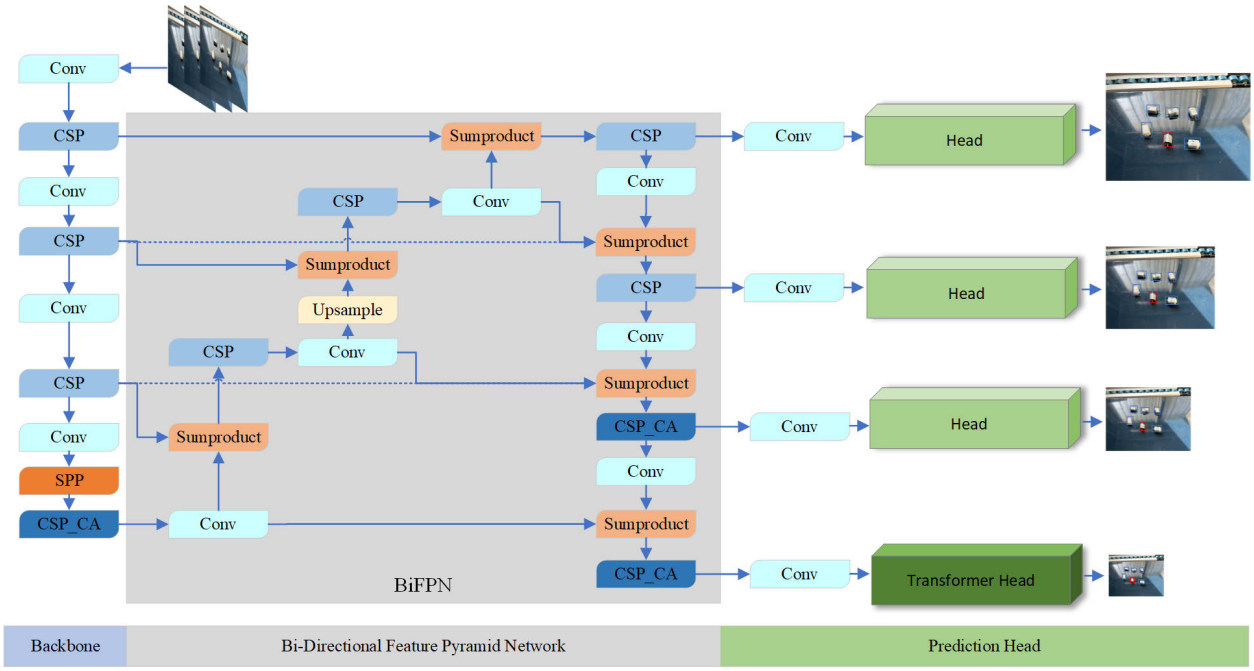


FIGURE 1. Structure of proposed network. In the inference stage, the input is a RGB image from camera, the output prediction is the primary picture with marker box. The CSP_CA represents the CSP module with Coordinate Attention.

fusion and feature map extraction, separately. Stacking convolutions and the CSP unit structure of CSPNet are exactly utilized to extract image features. This network is more lightweight than the DarkNet structure used by YOLOv3, enhancing the learning ability of the CNN while maintaining accuracy. The hierarchical features obtained through the backbone network are fused in the Neck part of the network. The FPN+PAN structure adopted by the model is not only a simple combination of multi-level features but also realizes path enhancement from low to high through upsampling, downsampling and residual connection. The method allows the model to fuse more levels of features to obtain a larger receptive field, and performs target prediction by combining detection heads of different resolutions, thereby achieving a good model prediction effect on targets of different scales.

III. METHODOLOGY

The proposed network structure is shown in Figure 1. Some deep-level features are extracted by adding the CSP unit with the CA module. The BiFPN is utilized to integrate the features, simplify a portion of the network structure, and pay close attention to obtain features at different levels. To locate and classify targets, the fused features are transmitted to the corresponding detectors in accordance with different resolutions.

A. COORDINATE ATTENTION

The images of industrial parts and information of included mechanical parts are usually accompanied by complex background environment. The YOLOv5 network uses stacking multiple CSP residual modules for feature extraction, which can continuously accumulate redundant information during network iteration and reduce the detection accuracy. In view of the confusion of targets during dense data detection, this paper optimizes the overall feature extraction ability of the

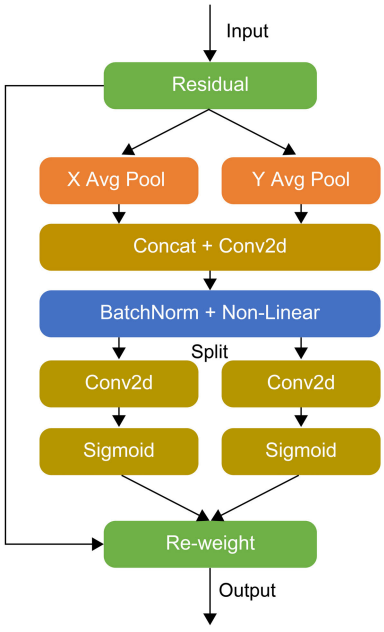


FIGURE 2. Structure of Coordinate Attention. It carries out average pooling in horizontal and vertical directions, then carries out transformation to encode spatial information, and finally fuses spatial information by weighting on the channel.

model, by embedding position information into the attention module after adding Coordinate Attention (CA) [31] into the CSP structure.

Attention mechanisms in computer vision, which aim to mimic the human visual system, can efficiently capture salient regions in complex scenes, making progress in multiple vision tasks. Through the attention mechanism, the input image features can be dynamically weighted. The SENet improves the recognition performance of the convolutional network by the feature extraction capability of the attention optimization model at the feature channel and spatial information level. But attention modules in methods such as SENet and CBAM [32] only consider internal channel information, ignoring the importance of location information. It is undeniable that the spatial structure of objects in vision is of great significance. Based on CBAM, coordinate attention is simplified, as shown in Figure 2. Give an input X, a pooling window of size (H, 1) or (1, W) is set along horizontal and vertical coordinates. By using the two parallel one-dimensional feature codes obtained from each channel, spatial coordinate information is integrated efficiently to acquire coordinate attention through the subsequent convolution structure to map the input features, so as to ensure that the network feature extraction ability is enhanced with less computational overhead, while obtaining more receptive field information.

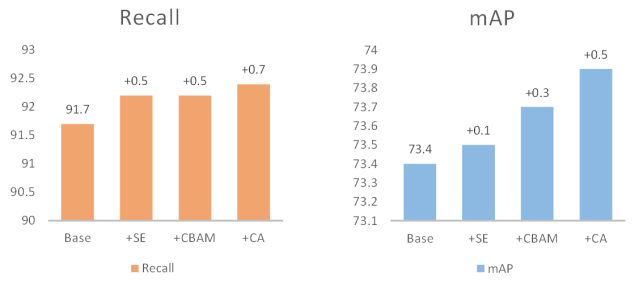


FIGURE 3. Comparison of the results in different attention methods. It inserts the SE blocks, CBAM blocks and CA blocks into the same position in the YOLOv5 model.

To demonstrate the advantages of the proposed method over other attention methods, experiments are conducted on industrial parts datasets under different improved structures. As shown in Figure 3, with a sufficient number of iterative training and computations, our network can achieve a 0.7% performance improvement in recall on the test set and a 0.5% improvement in average precision, which is significantly better than models with other attention mechanisms.

B. FEATURE FUSION WITH BiFPN

In one-stage object detection, the backbone network can extract more complex texture features with the increased number of layers, and the neck should fully integrate the features extracted from the backbone network. For the problem that the top-down FPN is limited by a single information flow, the PANet structure is employed in YOLOv5 presented in Figure 4(a). Bottom-up path aggregation is added based on the feature pyramid. The combination of upsampling and downsampling for multi-scale feature fusion can obtain deeper semantic information. However, the shallow features of the neck will be diluted, hindering the full combination of image features between deep layers and shallow layers. Considering many instances of small size in the defect detection dataset of industrial parts, and the difficulty in distinguishing features at the deep level, shallow features with in-depth features are combined by the BiFPN [33]. The attention computation enhances shallow feature information flow, making the model more biased towards small target samples in terms of assigning weights rather than direct summation, as in PANet.

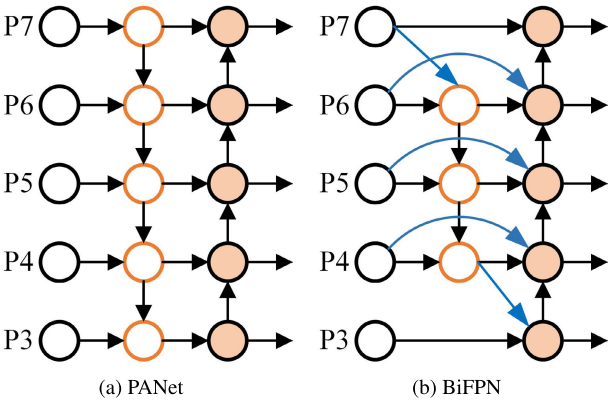


FIGURE 4. Comparison of structural differences between PANet and BiFPN. The left is the PANet in YOLOv5, the right is the BiFPN in our method.

As is depicted in Figure 4(b), the BiFPN is simplified based on PANet, using a weighting attention strategy, and adding additional residual connections to the same-level feature layer, which can fuse more layer-side features without increasing the amount of calculation, improve the feature fusion ability of the network, and effectively enhance the classification accuracy of difficult samples.

C. TRANSFORMER DETECTOR

Aiming at highly similar appearance of some defect categories of industrial parts samples, inspired by [34], a fine-grained object detector is designed in this paper by combing the advantages of the Transformer structure. Different from general fine-grained detection methods, the Transformer detector is utilized to enhance the model's ability to classify fine-grained categories, enabling end-to-end detection, and direct outputs of final detection results. As shown in Figure 5, the Transformer structure consists of two parts: the multi-head attention layer, and the feedforward neural network layers, which are connected by the residual structure. The Transformer encoder block increases the ability to capture different local image information, and can explore the feature representation potential through the self-attention mechanism to quickly distinguish similar samples. Combined with other prediction heads, this structure can alleviate the

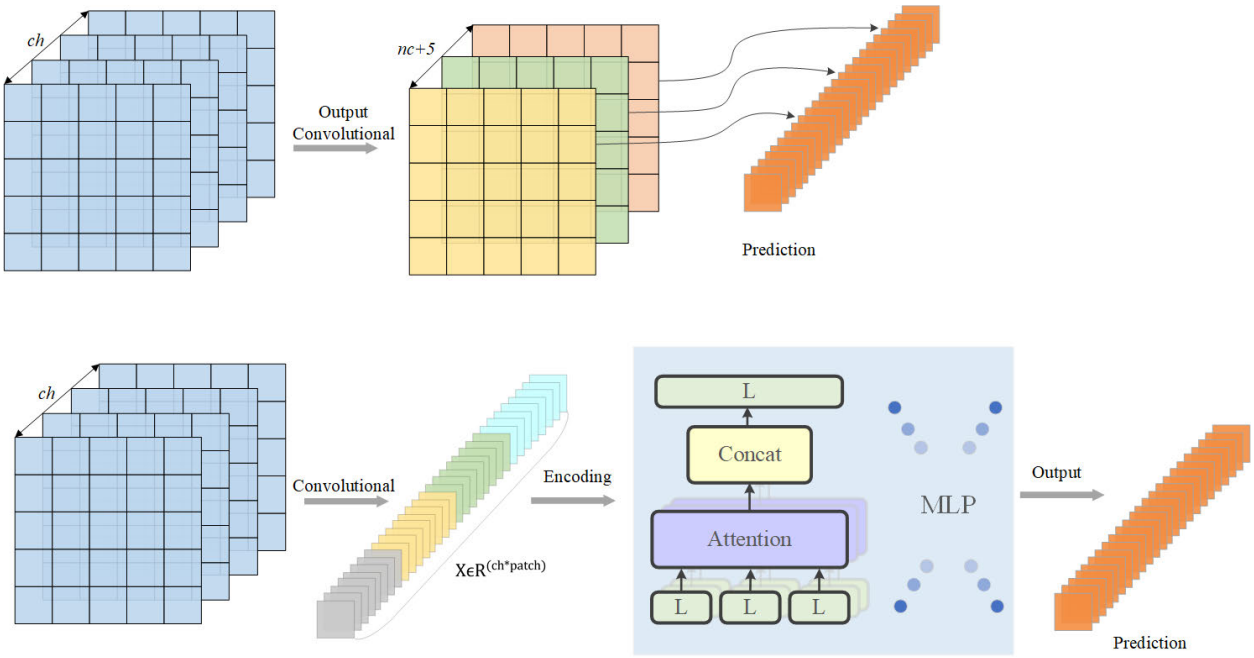


FIGURE 5. The Transformer Module. The top part is the standard convolutional predict head, the bottom part is the Transformer predict head consists of Multi-Head Attention, MLP and other modules. L represents the Linear layer.

TABLE 1. Results and params of the proposed model.

Structure	$\text{Recall}_{abnormal}$	Params(M)	FLOPs(G)	FPS
YOLOv5s	86.3	7.24	16.5	156
+CA	87.2(+0.9)	7.69(+0.45)	18.2(+1.7)	140
+BiFPN	88.9(+1.7)	7.78(+0.09)	18.5(+0.3)	135
+Transformer	91.6(+2.7)	10.95(+3.17)	27.9(+9.4)	95

adverse effects of drastic object scale changes. Despite costs of computation and memory caused by the additional detector, the performance of object detection has been greatly improved.

The improved network structure does not replace all CSP units with the improved module of fusion coordinate attention since the scale distribution of the dataset in this experiment is mostly small target samples, and pre-training weights cannot be effectively used due to the change in the structure. Adding attention to the beginning of the backbone network increases the difficulty of model training, which may result in unstable final detection performance. Therefore, CSP units are only replaced in some areas to avoid unnecessary computational overhead and ensure the robustness of the model. The predicted recall rate and parameters of the model are summarized in Table 1.

FIGURE 6. The data samples of the Industrial Part Defect Dataset. The normal target is the normal status of the industrial part, the others represent 4 abnormal status.

IV. EXPERIMENT AND ANALYSIS

A. DATASET

Given large variety of industrial parts and different types of transportation products on assembly lines, there is currently no unified public dataset. Aiming at the detection of micro-motor defects in industrial production, this paper collected videos and images of the same kind of industrial motors on assembly lines in different environments, and then organized and performed data augmentation to produce a motor defect

detection dataset for assembly line operation scenarios. The dataset contains 1400 images labeled and exported using the EasyData labeling platform of Baidu Smart Cloud. As introduced in Figure 6, labeling categories include normal, dirty, structural distortion, main body deformation, and incomplete. A total of 8613 labeling boxes are obtained, including 5837 normal labels, 820 dirty labels, 778 twist labels, 531 deformative labels, and 647 incomplete labels. Due to safety and industrial production requirements, most of the images are taken from a distance above the assembly line, coupled with the target industrial motors in small size in the experiment, thereby generating tiny targets collected in the dataset and many dense distributions of targets.

By shuffling the dataset order, 80% of the data is selected randomly as the training set and 20% as the validation set. In addition, different strategies are applied for data augmentation in both training and inference stages to reinforce the model training accuracy. The Mosaic method is randomly used during model training, including affine transformation, random rotation, translation, scaling, cropping, flipping, and other data augmentation methods. For inference verification, only scaling and normalization are utilized.

B. EVALUATION METRICS

The evaluation indicators in this paper contain Recall, Precision, and Mean Average Precision (mAP), commonly used in object detection. For abnormal sample detection, the recall rate of abnormal samples is an important indicator for method evaluation. The recall rate refers to the proportion of all targets predicted by the model that is correctly predicted. As a widely used evaluation index in the field of defect detection, the recall is related to T_P and F_P.

The evaluation indicators in this paper contain Recall, Precision, and Mean Average Precision (mAP), commonly used in object detection. For abnormal sample detection, the recall rate of abnormal samples is an important indicator for method evaluation. The recall rate refers to the proportion of all targets predicted by the model that is correctly predicted.

$$Recall = \frac{T_P}{T_P + F_N} \tag{1}$$

The prediction accuracy represents the proportion of all targets predicted by the model that is correctly expected.

$$Precision = \frac{T_p}{T_P + F_P} \tag{2}$$

The mAP is currently the most popular evaluation metric in object detection, and its calculated value involves several related concepts. The intersection and the union ratio, IoU, measures the degree of overlap between the two regions and is the ratio of the overlapping area of the two regions to the total area. As is shown in Figure 7, the IoU of two rectangular boxes is the ratio of the intersection area to the combined area.

$$IoU = \frac{S2}{S1 + S2 + S3} \tag{3}$$

FIGURE 7. IoU calculation diagram. The Ground truth box is the true label of the target like the S1, the prediction box is the output prediction from model like the S3, the overlapping area is the S2, and the IoU is calculated by the S1, S2 and S3.

The PR (Precision-Recall) curve can reflect the performance of an algorithm, by setting a calculation threshold θ as the threshold for determining whether the prediction result is a positive or negative sample. IoU is often used in object detection to determine the prediction result. For example, first set IoU ≥ 0.5 as the same target, then it is determined that the prediction that meets the condition is true, otherwise, it is false. Then calculate the Precision and Recall of the data set, then increase the threshold θ in a decreasing manner, and record the Precision and Recall corresponding to the respective thresholds, each threshold θ corresponds to a (Precision, Recall) point, and connecting these points is PR curve.

Average Precision (AP): An evaluation index reconciles precision and recall's contradictory variables in object detection. The recall is the horizontal axis, the precision is the vertical axis, and the PR curve encloses the area of the irregular graph. Since the integral calculation is relatively tricky, approximate interpolation calculation is adopted.

$$AP = \frac{1}{n}\sum_{i=1}^{n-1} P_{interp}(r) \tag{4}$$

where $P_{interp}(r)$ is the larger value of the accuracy in the r position and the r next position. Mean Average Precision (mAP): to improve the comprehensiveness of the calculation accuracy, 100 points were sampled on the PR curve for calculation. And the threshold of IoU is adjusted from a fixed value of 0.5 to the value of AP calculated every 0.5 in the interval of 0.5 - 0.95, and the average of all results is taken as the final result.

$$mAP = \frac{1}{m}\sum_{i=1}^{m} AP_i \tag{5}$$

C. ABLATION STUDY

To verify the generalization of the model, the performance of the proposed model is compared with other models like RetinaNet, EfficientDet-D0, and YOLOv5s in the defect detection dataset of industrial parts. Pre-trained weights on the ImageNet dataset are used for all models to complete transfer learning. SGD is adopted for optimization during the training process. The number of training images per batch is

TABLE 2. Industrial parts defect detection experiment comparison.

Model	Backbone	Input Size	AP(%)	$AP_{50}(\%)$	$AP_{75}(\%)$	Time(ms)
YOLOv5s	CSPDarknet	640	73.4	91.3	86.6	6.4
EfficientDet-D0	Efficientnet	640	71.7	89.7	84.1	10.2
RetinaNet	ResNet50	640	74.2	91.7	86.9	17.9
YOLOR	BottleneckCSP	640	75.1	92.8	88.0	31.4
CenterNet2	ResNet50	640	74.4	92.3	87.1	29.6
SwinTransformer	-	640	76.3	94.4	90.5	67.1
Ours	CSP_CA	640	75.6	93.6	89.2	10.5

TABLE 3. Detection results of different proposed structures on an industrial part defect dataset. Recall(abnormal) represents the recall rate except normal target.

CA	BiFPN	Transformer	Recall(%)	$Recall_{abnormal}(\%)$	AP(%)	Time(ms)
			91.7	86.3	73.4	6.4
✓			92.4	87.2	73.9	7.1
	✓		93.3	88.6	74.1	6.5
✓	✓		93.5	88.9	74.4	7.4
		✓	93.7	89.5	75.0	9.3
✓		✓	93.9	90.4	75.2	9.9
	✓	✓	94.3	91.2	75.5	10.1
✓	✓	✓	94.5	91.6	75.6	10.5

slightly different according to different networks. Based on experience, the initial learning rate is set to batch size * 0.001, the learning rate decay is performed once every 60 iterations, a total of 200 epochs, and the decay coefficient is 0.1 to converge the model parameters further. The input image size of the selected model is 640*640 RGB image. The model is evaluated using the validation set after each round of training, and the validation curve is shown in Figure 8.

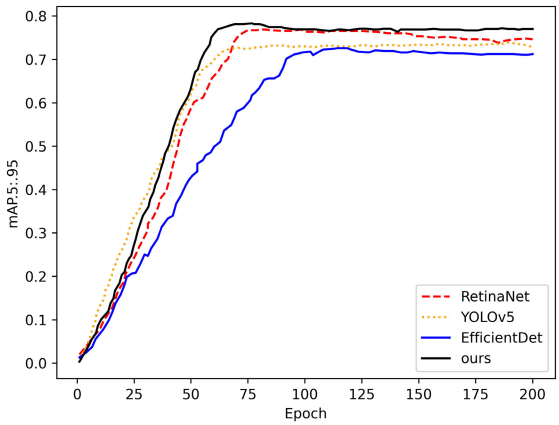


FIGURE 8. Validation curve of epochs. It compares the validation mAP from RetinaNet, YOLOv5, EfficientDet-D0 and our method in the training stage.

As can be seen from Figure 8, the proposed model achieves higher detection accuracy than YOLOv5s, and the mAP reaches 0.756, which is 2.2 percentages points higher than the original network. It can be seen from Table 2 that the proposed method achieves almost the highest detection accuracy among the same type of methods, reaching 93.6%, and the detection speed is also at a high level. The inference speed on

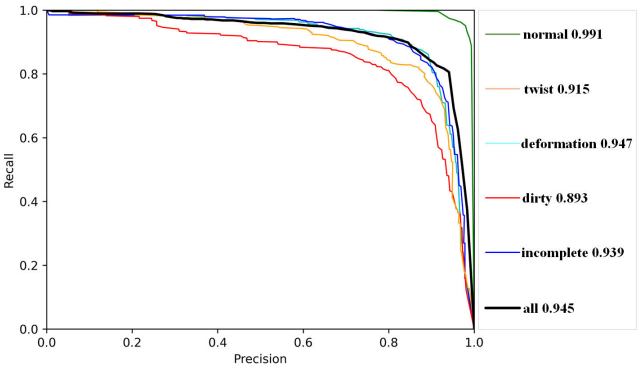
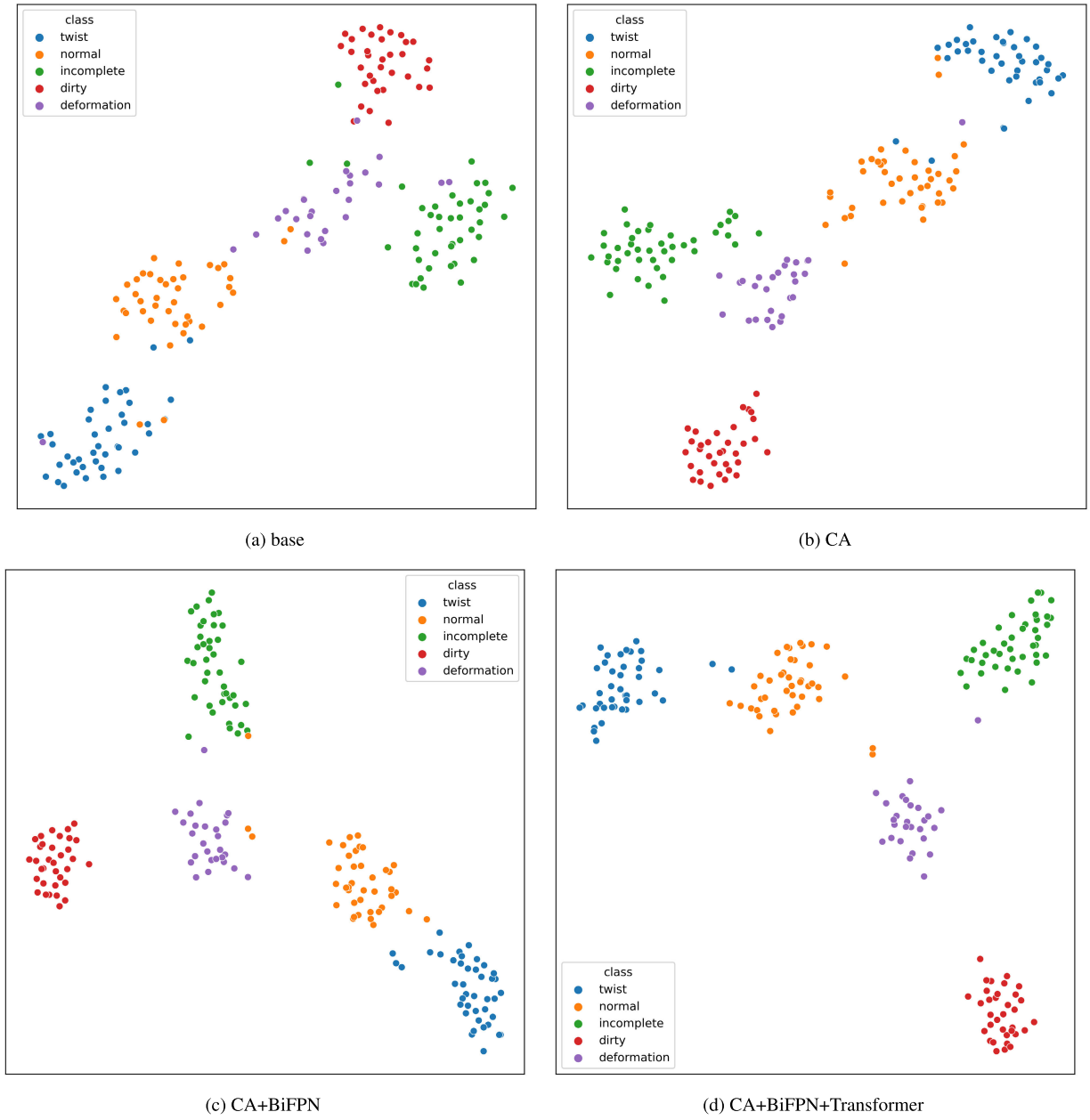


FIGURE 9. PR curve of the proposed model in industrial part defect dataset.

the A100 graphics card reaches 95FPS, which still meets the needs of real-time detection. The experiments demonstrated that our model remains competitive on the dataset.

To verify the optimization effect of each proposed module in the network, ablation experiments are carried out according to the proposed method. The experimental results are summarized in Table 3. The $Recall_{abnormal}$ is the primary evaluation indicator of the abnormal detection of industrial parts. After adding the coordinate attention module, the average precision of the model is increased by 0.5%. After using the bidirectional multi-scale fusion module for feature integration, the abnormal recall rate of the detection model is increased by 2.3%. It is concluded that the prediction accuracy of the method for small target samples is significantly improved. The addition of the Transformer detector guarantees great enhancement of the recall rate and precision. Figure 9 shows the precision-recall curve of the detection performance of proposed model. The detection speed decreases due to the



(a) base (b) CA

(c) CA+BiFPN (d) CA+BiFPN+Transformer

FIGURE 10. Feature dimensionality reduction visualization in t-SNE. The feature is from the feature maps of the penultimate layer of network, and reduced to 2 dimensions by PCA. Different colored dots represent different categories.

increased number of parameters and computation brought by its structure. The results show the proposed detection model is superior to the primary YOLOv5.

D. ANALYSIS

Figure 10 presents a comparison of the prototype distribution of classification features learned by both original and proposed models, indicating that the model with the bidirectional multi-scale fusion module still faces a small amount of sample confusion after network fine-tuning, but its classification interval is more apparent. The model added to the Transformer detector clearly distinguishes the vast majority of samples, reduces the overlap between categories, and realizes more balanced overall spacing of features, proving that the proposed method can improve the representation ability of the feature space for effective object detection.

Specifically, to compare the proposed network results more intuitively, some pictures in the test dataset and real pictures were selected for testing. For more obvious comparison results, the two networks' confidence thresholds were set to 0.45. The non-maximum suppression IoU threshold is set to 0.3.

Figure 11 describes the detection results of the YOLOv5s model and proposed model respectively on the left and

(a) small

(b) dense

FIGURE 11. Prediction results with marker box. The left side of the predict picture is from YOLOv5, and the right side is from proposed model. The part (a) shows the comparing results in small target detection situation. The part (b) shows the comparing detection results in dense situation.

right sides. In Figure 11(a), owing to long distance from the detection target to the shooting acquisition device, the detected targets tend to be tiny overall, and the confidence is lower with some false detections. The fact is that small target objects can be detected more accurately on the right side. In Figure 11(b), dense targets make some of the prediction frames on the left inaccurate and undetected, while the detection results on the right are improved.

Ablation experiments are carried out to verify the efficiency of the proposed module in the actual production environment. We set up control groups in the factory based on different environments. Each control group contains 20 batches of samples from the abnormal category, with four different abnormal samples in each batch. On a assembly line, samples from the same batch were photographed in different environments. The camera is situated between 65 and 85 cm away from the object in the normal control group. There is around 170 lx of indoor illumination, and the camera is brand-new. The samples were placed in two groups of experimental settings that were separated from the control groups by distances of 100 cm and 120 cm. The illuminance comparison group was set up with two distinct illuminance environments, namely the low-light group with an illuminance of approximately 100 lx and a high-light group with an illuminance of about 220 lx. A dirty-camera control group was established to shoot with the experiment using the same model camera that had been in use in the factory for about 14 months. While maintaining the same environment, all batches were shot in five shots with fine-tuning of the shooting angle, and total of 400 samples were collected. By using data enhancement methods including horizontal flip, vertical flip, and random cropping, the dataset was enlarged. With a total of 1600 samples, they were then summarized and sorted into an industrial parts environmental comparison dataset.

The prediction results of the proposed model on this dataset are shown in Table 4. It can be seen that the prediction recall rate of the model in different environments has been affected to different degrees. Among them, the Twist samples are more significantly affected when the camera is far away and dirty, with a maximum drop is about 5 percentage points. In low

TABLE 4. Prediction results in different environments.

Structure	True Positive				Recall(%)
	Twist	Incomplete	Deformative	Dirty	Total
normal	341	377	379	363	91.3
distance-100cm	328	374	378	356	89.8
distance-120cm	320	372	372	354	88.7
low-light	332	367	375	336	88.1
high-light	335	379	372	358	90.3
dirty-camera	324	365	373	341	87.8

light conditions, dirty class samples are more affected, and the recall rate decreases by 6.7 percentage points. The recall rate of the rest of categories is slightly influenced by the environment. Additionally, it was discovered throughout the experiment that the samples from the Twist category and the Incomplete category were marginally impacted by the shooting angle. According to the comparative experiments mentioned above, it can be found that the proposed method will slightly reduce the recall rate of abnormal samples when the illumination and camera height of the real production environment change slightly, but it can still meet the detection requirements.

The proposed model is suitable for use with embedded devices. After compiling to onnx, we migrate the model to NVIDIA Jetson NX and build a detection system based on it. Due to the limited computing power of the device, the detection speed on Jetson NX after porting is about 35FPS. When the detection system uses monitors to output the detection videos, the detection speed of the model decreases because the videos take up part of the computation, and it declines to 31FPS in our experimental environment. The above data are measured under the condition that the detection accuracy is unaffected and the model is transplanted without quantification. The model can be quantized in order to reduce the number of model parameters and calculations and speed up the inference for embedded devices to achieve real-time detection. The quantization operation will result in some recall loss that is related to the model compression rate.

V. CONCLUSION

In this study, we propose an end-to-end lightweight defect detection model for industrial parts based on improved YOLOv5. The detector can achieve excellent detection accuracy and real-time detection on edge computing device. Our contributions mainly concentrate on three aspects: applying the coordinate attention to module for feature extraction to improve the detection performance of the model, optimizing the model hierarchy through the BiFPN to reduce the false detection rate and missed detection of small target samples, and adding the Transformer detector to increase the recognition accuracy of difficult samples. The experimental results demonstrate that the algorithm proposed in this article improves the performance of the defect detection algorithm based on industrial parts under the premise of real-time detection and can help improve the yield in industrial production, transportation, and other scenarios. Currently, the algorithm still has a slight shortage of detecting the defects of parts with occlusion under a fixed shooting angle. Future research will further adjust the structure and determine how to improve the recognition accuracy through multi-angle collaborative detection to achieve better detection performance.

A. ABBREVIATIONS

AP	Averaged AP at IoUs from 0.5 to 0.95 with an interval of 0.05
AP_{50}	AP at IoU threshold 0.5
AP_{75}	AP at IoU threshold 0.75
BiFPN	Bi-directional feature pyramid network
CA	Coordinate Attention
CBAM	Convolutional block attention model
end-to-end	The input is the original data, and the output is the final result
FLOPs	Floating-point operations per second
FPN	Feature pyramid network
IoU	Intersection over union
lx	Lux, the unit of illumination.
$Recall_{abnormal}$	recall rate of abnormal samples
SSD	Single Shot multibox Detector
YOLO	You Only Look Once

REFERENCES

[1] H. Wang, J. Wang, G. Zhang, X. Ouyang, and F. Luo, "Improved FPN's mask R-CNN for industrial surface defect detection," *Manuf. Automat.*, vol. 42, no. 12, pp. 35–40 and 97, Dec. 2020.

[2] Y. Chen, Y. Ding, F. Zhao, E. Zhang, Z. Wu, and L. Shao, "Surface defect detection methods for industrial products: A review," *Appl. Sci.*, vol. 11, no. 16, p. 7657, Aug. 2021.

[3] H.-Y. Lee and T.-E. Lee, "Scheduling single-armed cluster tools with reentrant wafer flows," *IEEE Trans. Semicond. Manuf.*, vol. 19, no. 2, pp. 226–240, May 2006.

[4] D. V. Slavov and V. D. Hristov, "3D machine vision system for defect inspection and robot guidance," in *Proc. 57th Int. Sci. Conf. Inf., Commun. Energy Syst. Technol. (ICEST)*, Jun. 2022, pp. 1–5.

[5] M. Foumani, M. Y. Ibrahim, and I. Gunawan, "Scheduling dual gripper robotic cells with a hub machine," in *Proc. IEEE Int. Symp. Ind. Electron.*, May 2013, pp. 1–6.

[6] T. Czimmermann, G. Ciuti, M. Milazzo, M. Chiurazzi, S. Roccella, C. M. Oddo, and P. Dario, "Visual-based defect detection and classification approaches for industrial applications—A SURVEY," *Sensors*, vol. 20, no. 5, p. 1459, Mar. 2020.

[7] H. Chang, J. Gou, and X. Li, "Application of faster R-CNN in image defect detection of industrial CT," *J. Image Graph.*, vol. 23, no. 7, pp. 1061–1071, 2018.

[8] X. Yue, Q. Wang, L. He, Y. Li, and D. Tang, "Research on tiny target detection technology of fabric defects based on improved Yolo," *Appl. Sci.*, vol. 12, no. 13, p. 6823, Jul. 2022.

[9] Z. Li, X. Tian, X. Liu, Y. Liu, and X. Shi, "A two-stage industrial defect detection framework based on improved-YOLOv5 and optimized-inception-ResNetV2 models," *Appl. Sci.*, vol. 12, no. 2, p. 834, Jan. 2022.

[10] J. Božic, D. Tabernik, and D. Skocaj, "Mixed supervision for surface-defect detection: From weakly to fully supervised learning," *Comput. Ind.*, vol. 129, Aug. 2021, Art. no. 103459.

[11] Q. Luo, X. Fang, L. Liu, C. Yang, and Y. Sun, "Automated visual defect detection for flat steel surface: A survey," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 3, pp. 626–644, Mar. 2020.

[12] J. Yang, S. Li, Z. Wang, and G. Yang, "Real-time tiny part defect detection system in manufacturing using deep learning," *IEEE Access*, vol. 7, pp. 89278–89291, 2019.

[13] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[14] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 17–24.
[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.
[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[18] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.
[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[27] H T. Lu and Q. C. Zhang, "Applications of deep convolutional neural network in computer vision," *J. Data Acquisition Process.*, vol. 31, no. 1, pp. 1–17, 2016.
[28] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 390–391.
[29] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[30] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
[31] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.
[32] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[33] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.
[34] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.

HAI FENG LE received the bachelor's degree from the School of College of Urban Rail Transit and Logistics, Beijing Union University, in 2019. He is currently pursuing the graduate degree with the School of Robotics, Beijing Union University. His interests include deep learning and applications and computer graphics.

LU JIA ZHANG was born in Beijing, China, in 1996. She received the bachelor's degree in computer science and technology from Beijing Union University's Smart City College, in 2019. She is currently pursuing the graduate degree with the School of Robotics, Beijing Union University. Her research interests include image recognition and deep learning and applications.

YAN XIA LIU received the Ph.D. degree from the School of Automation and Electrical Engineering, University of Science and Technology, Beijing, in 2013. She is a Professor with the College of Urban Rail Transit and Logistics, Beijing Union University. Her current research interests include pattern recognition, computer vision, deep learning, and intelligent instruments.