ACR-Net: Attention Integrated and Cross-Spatial Feature Fused Rotation Network for Tubular Solder Joint Detection

Chenlin Zhou, Daheng Li, Peng Wang, *Member, IEEE*, Jia Sun, Yikun Huang, and Wanyi Li

Abstract— **Tubular solder joint detection is an important and challenging issue in the industry, due to the illegible objects, rarely collected datasets and requiring high-precision and real-time performance for positioning and angle estimation. In this article, we propose an Attention integrated and Cross-spatial feature fused Rotation Network (ACR-Net) for tubular solder joint detection, which consists of a new feature extraction network named ECA-CSPDarknet44, a cross-spatial feature fusion network (CFFN), and a bin-based rotation detection network (BRDN). The proposed network can efficiently detect oriented tubular solder joints with high-precision and real-time performance. ECA-CSPDarknet44 with attention mechanism was presented, which can adaptively guide the network to learn important features of tubular solder joints, significantly improving the ability of feature extraction. By integrating multiscale global features and multiscale local region features, CFFN can enhance the network's ability to express the characteristics of tubular solder joints. Meanwhile, BRDN is proposed for oriented bounding box regression through a decoupling approach, which regresses target parameters efficiently and accurately with little increasing of model complexity. Finally, we establish a tubular solder joint dataset and conduct sufficient experiments to verify the effectiveness of our method. Our proposed ACR-Net achieves 98.8% mAP with 31.3 frames per second (FPSs) on the dataset, meeting the high-precision and real-time requirements of industrial systems.**

Index Terms— **Deep learning, defect detection, object detection, tubular solder joint detection.**

I. Introduction

Tubular solder joint detection mainly comes from industrial defect detection or quality inspection. In the industrial production line, the manipulator system detects the quality of tubular solder joints that are usually at the junction or the end of the metal pipe in the compressor of the refrigerator. During the inspection, the manipulator approaches targets by obtaining the center point coordinates of tubular solder joints. Then, a suction leak detector, which is usually installed at the end of the manipulator, is used to check the quality of the tubular solder joint. In order to reduce or even avoid collisions with a metal pipe, the suction leak detector needs to wrap a tubular solder joint along the metal pipe. Therefore, tubular solder joint detection consists of both the center point (x, y) positioning and angle θ estimation of tubular solder joint in the perception module of industrial manipulator systems.

Most studies related to solder joint detection are about printed circuit board (PCB) solder joint [1]–[6] and metal surface weld [7]. The method on tubular solder joint detection is rarely proposed. PCB solder joint detection is essentially aimed at solving a classification task of different defect types. In metal surface weld detection, because the background of the weld is usually structured and simple, it is easy to identify or locate by hand-designed methods. Comparatively, tubular solder joint detection is a more challenging issue, and the targets are small and usually account for less than 1/5000 of the whole image. The background is complex and low contrast, and objects are hard to identify or locate. The detection of manipulator systems in industrial production lines requires real-time and high-precision performance. Furthermore, tubular solder joint detection includes both the center point positioning and angle estimation, which is unnecessary in the other solder joint detections. All in all, these factors bring great difficulties to tubular solder joint detection.

Deep learning methods have achieved good performances on image classification [8] and object detection tasks [9] because of theirs powerful feature extraction and feature expression capability. They have been also widely used in industry, such as gear pitting detection [10], quality inspection [11], and defect detection [12]–[14]. However, there are two main problems in applying deep learning methods to tubular solder joints detection. On the one hand, from the point of general object detection, it is difficult to directly detect the target with the form of (x, y, θ), which is not conducive to network training, optimization, and evaluation. As shown in Fig. 1(b), in order to facilitate network processing, we convert this problem into oriented object detection with the regression form of (x, y, w, h, θ), which includes the center point and

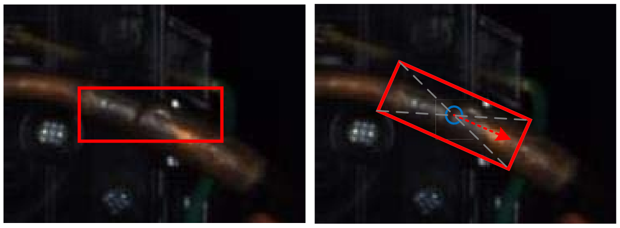
Manuscript received March 20, 2021; revised June 21, 2021; accepted June 26, 2021. Date of publication July 5, 2021; date of current version July 21, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 91748131, Grant 62006229, and Grant 61771471; in part by the Strategic Priority Research Program of Chinese Academy of Science under Grant XDB32050100; and in part by the Beijing Municipal Natural Science Foundation under Grant 4204113. The Associate Editor coordinating the review process was Anirban Mukherjee. *(Corresponding author: Peng Wang.)*

Chenlin Zhou and Daheng Li are with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhouchenlin2019@ia.ac.cn).

Peng Wang is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China (e-mail: peng_wang@ia.ac.cn).

Jia Sun, Yikun Huang, and Wanyi Li are with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jia.sun@ia.ac.cn).

Digital Object Identifier 10.1109/TIM.2021.3094837



(a) Detection result based on horizontal object detection method (b) Detection result based on oriented bounding box in our method

Fig. 1. Results of tubular solder joint detection. (a) Detection result of horizontal object detection method with the horizontal bounding box, lacking angle estimation. (b) Detection result based on our proposed ACR-Net, containing both the positioning and angle estimation of tubular solder joint.

orientation of tubular solder joint. On the other hand, oriented object detectors mainly aim at aerial images and text detection and perform well on public datasets. In specific application scenarios, such as tubular solder joint detection, these oriented object detectors cannot meet the high-precision and real-time requirements of industrial systems due to redundant and unreasonable structural design caused by obvious differences in target characteristics. To deal with these problems, we propose an Attention integrated and Cross-spatial feature fused Rotation Network (ACR-Net) for tubular solder joint detection, which consists of a feature extraction network ECA-CSPDarknet44, a cross-spatial feature fusion network (CFFN), and a bin-based rotation detection network (BRDN). It can efficiently detect oriented tubular solder joints with high-precision and real-time performance. The contributions of our work are shown as follows.

1) We propose ACR-Net for tubular solder joint detection, which can efficiently detect oriented tubular solder joints with state-of-the-art performance.
2) A new feature extraction network ECA-CSPDarknet44 is proposed for tubular solder joint detection. It can adaptively guide the network to learn the important features of solder joints, significantly improving the ability of feature extraction for tubular solder joints.
3) We design a CFFN that integrates global features of multiscale convolutional layers and multiscale local region features on the same convolutional layer, improving the ability to express the characteristics of tubular solder joints.
4) BRDN is presented for oriented bounding box regression through a decoupling approach. It regresses target parameters efficiently and accurately with little an increase in model complexity.
5) We establish a tubular solder joint dataset and conduct sufficient experiments to verify the effectiveness of our method. ACR-Net achieves 98.8% mAP with 31.3 frames per second (FPSs) on the dataset, meeting the high-precision and real-time requirements of industrial systems.

II. Related Work

The works related to tubular solder joint detection can be summarized into the following three parts: horizontal object detection, oriented object detection, and solder joint detection.

A. Horizontal Object Detection

Many high-performance object detectors have been proposed to detect general objects with horizontal bounding boxes. These horizontal object detection models can be divided into anchor-based detectors and anchor-free detectors, according to whether or not the detectors use the preset anchors. Anchor-based detectors can be categorized into one-stage detectors and two-stage detectors. Two-stage detectors generate region proposals first and then predict the precise location of targets and the corresponding category labels. The most representative two-stage models are R-CNN [9] series, including fast R-CNN [15], faster RCNN [16], R-FCN [17], and Libra R-CNN [18]. One-stage detectors directly use a convolutional neural network (CNN) for horizontal bounding box regression and object classification without the process of region proposals. The most representative one-stage models include YOLO [5], [19], [20], SSD [21], and RetinaNet [22]. Anchor-free detectors include CenterNet [23], CornerNet [24], FCOS [25], RepPoints [26], and so on. Two-stage detectors are seldom used in industrial scenes with the high real-time requirement because they are more accurate but slower than one-stage detectors in general. These horizontal object detectors have achieved great performance on general objects, but these methods cannot be directly used to solve tubular solder joint detection, due to lacking the mechanism for detecting the orientation of tubular solder joints.

B. Oriented Object Detection

Aerial images and text are the main application scenarios of oriented object detection. Recent advances in oriented object detection are mainly driven by adaptation of classical horizontal object detection methods [27].

1) Object Detection in Aerial Images: Due to the complexity of remote sensing images and a large number of small, cluttered, and rotated objects, two-stage oriented detectors are still dominant for their robustness [28]. Rotated RPN is exploited in [29] and [30], and it needs more calculations and runtime because of involving lots of rotated anchors. The RoI transformer [7] learns from the horizontal anchor to obtain a rotating region of interest (ROI), reducing the amount of calculation, and extracts features based on the rotating ROI, on which the rotated bounding box regression is implemented. R3Det [28] uses a progressive regression approach from coarse to fine granularity. It predicts coarse rotation box based on horizontal anchor boxes rapidly and then performs oriented object detection based on coarse rotation box by a feature refinement module. RSdet [31] proposes a modulated rotation loss to dismiss the loss discontinuity, which is caused by resulted from the inherent periodicity of angles and the associated sudden exchange of width and height. ICN [32] proposes a new method consisting of a novel joint image cascade and feature pyramid network with multisize convolution kernels to extract multiscale strong and weak semantic features. In [27], a novel method of rotating bounding box representation based on a gliding vertex on the horizontal bounding box is introduced to describe multioriented objects more accurately and avoid confusion issues.

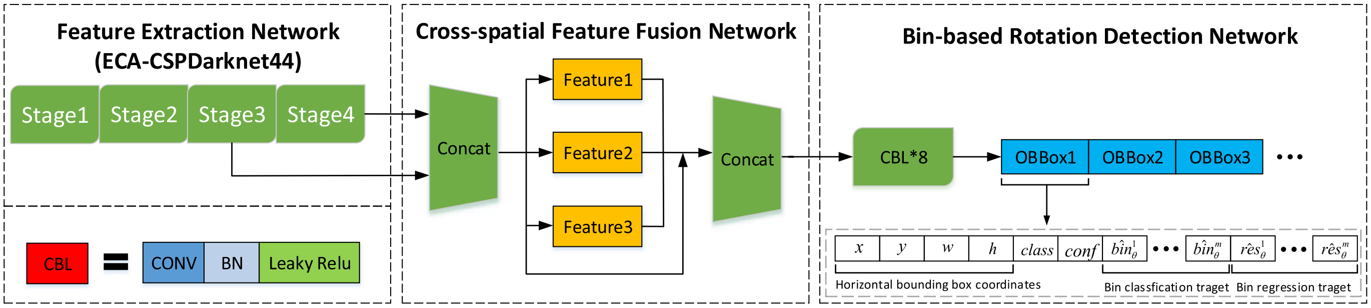


Fig. 2. Pipeline of our proposed ACR-Net. It is an end-to-end network and consists of a feature extraction network named ECA-CSPDarknet44, a CFFN, and a BRDN. The prediction of ACR-Net is oriented bounding box: $(B_x^{\text{pre}}, B_y^{\text{pre}}, B_w^{\text{pre}}, B_h^{\text{pre}}, B_\theta^{\text{pre}})$. B_θ^{pre} can be decoded from the bin classification target $\hat{\text{bin}}_\theta : [\hat{\text{bin}}_\theta^1, \hat{\text{bin}}_\theta^2, \ldots, \hat{\text{bin}}_\theta^m]$ and the residual regression target $\hat{\text{res}}_\theta : [\hat{\text{res}}_\theta^1, \hat{\text{res}}_\theta^2, \ldots, \hat{\text{res}}_\theta^m]$ by formula (10).

2) Oriented Text Detection: Most methods directly use rotated bounding box or quadrangle representation to detect entire targets in text detection. RRPN [33] employs rotated RPN in the framework of faster R-CNN [16] to generate rotated proposals and further performs rotated bounding box regression. TextBoxes++ [34] adopts vertex regression on SSD. RRD [35] implements classification and bounding box regression on rotation-invariant and rotation-sensitive features, extracted by two network branches of different designs, respectively. DMPNet [36] detects text by a prior quadrilateral sliding window, which significantly improves the recall rate. FOTS [37] achieves simultaneous detection and recognition by a unified end-to-end network, sharing computation and visual information among the two complementary tasks. Wang *et al.* [38] propose an end-to-end text spotting method by a set of points on the boundary of each text instance, which can read the text of arbitrary shapes.

These general oriented object detectors mainly aim at aerial images and text detection. Due to obvious differences in target characteristics and application requirements, these detectors have a redundant and unreasonable structural design in specific application scenarios, such as tubular solder joint detection. These methods cannot meet high-precision and real-time requirements in industrial tubular solder joint detection.

C. Solder Joint Detection

Most studies related to solder joint detection are about PCB solder joint and metal surface weld. The majority of methods on PCB solder joint detection aim at solving a classification task of different defect types, which represent the different types of connection between component and solder joint on the PCB after mounting. The majority of methods [7] on metal surface weld detection are used to locate weld. Then, some special instruments, such as an ultrasonic flaw detector, are used to inspect the quality of the weld. The background of the metal surface weld is usually structured and simple, and it is easy to identify or locate by hand-designed methods. Tubular solder joint detection includes both the center point positioning and angle estimation of the tubular solder joint. However, the methods on PCB solder joint and metal surface weld detection lack the mechanism for detecting the direction of solder joints. Meanwhile, it is difficult to transfer these methods because they follow task-specific modeling on the whole. In addition, tubular solder joints are small and illegible, and the background is complex and low contrast. Traditional solder joint detection methods are difficult to solve these problems; therefore, advanced detection algorithms are urgently needed for tubular solder joint detection.

III. Proposed Method

The proposed ACR-Net is a one-stage anchor-based oriented object detector for tubular solder joint detection. In this section, the details of our method are discussed. As shown in Fig. 2, ACR-Net contains a feature extraction network named ECA-CSPDarknet44, a CFFN, and a BRDN. Finally, a new evaluation index for tubular solder joint detection is discussed.

A. Feature Extraction Network

ECA-CSPDarknet44 is proposed for feature extraction of tubular solder joints, which combines CSPDarknet53 [5] and channel attention module ECA [39].

CSPDarknet53 is composed of cross-stage partial network (CSPNet) [40] and Darknet53 [20]. Darknet53 is a widely used feature extraction module. It has five stages, and each stage consists of several repeated residual modules. At the same time, each stage represents one downsampling. After five downsamplings, the size of the output feature map of the backbone is 1/32 of the input. In each stage, the number of repeated residual modules n is 1, 2, 8, 8, and 4, respectively. CSPNet is used to reduce computation and promote the learning capability of a CNN when ensuring equivalent or even superior accuracy by integrating feature maps from the beginning and the end of each stage. Fig. 3(a) shows the structure of a CSPDarknet53 stage. In order to make it easier to understand, it also can be considered to be composed of the downsampling module, the body module, and the channel concat module. In the downsampling module, it is achieved by setting the stride of convolution as 2. After passing through this convolutional layer, the size of the feature map will be reduced to half. The input feature maps are separated into two 2 branches in the body module. Branch 1 is a partial transition layer, which is directly linked to the channel concat module through a transition layer. Branch 2 is a partial dense block, mainly composed of a Darknet53 stage. In the channel concat module, both the end of the partial dense block and partial

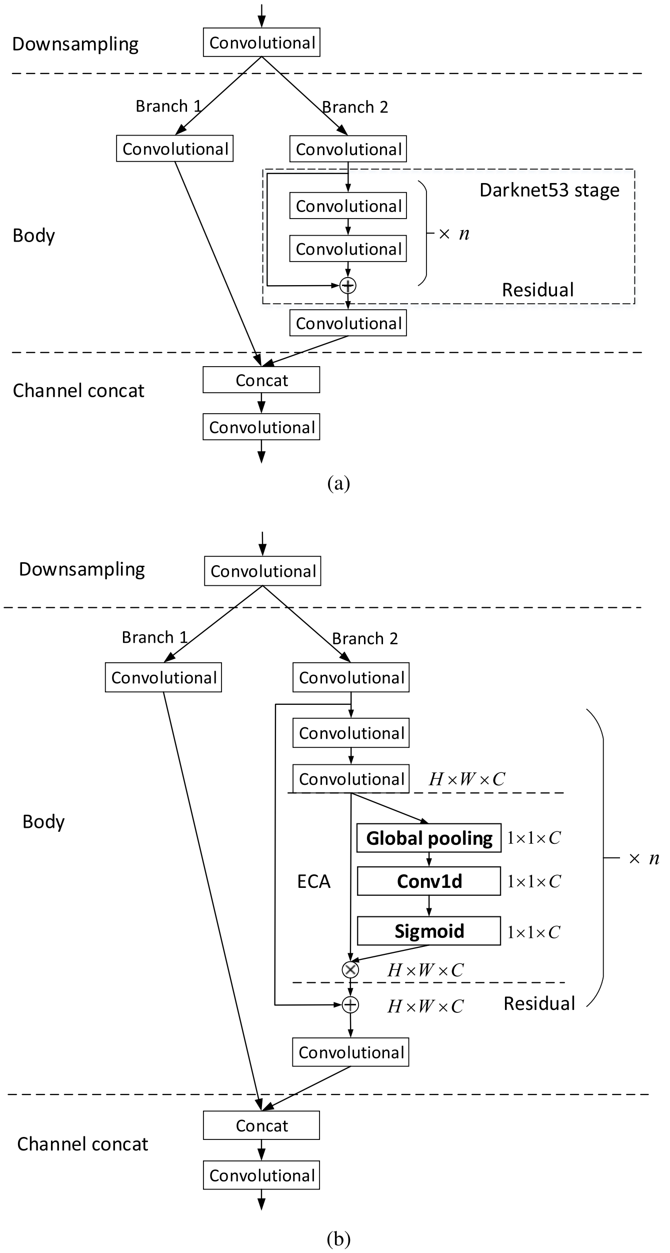


Fig. 3. Stage difference between CSPDarknet53 and proposed ECA-CSPDarknet44 in the feature extraction network. (a) Structure of a CSPDarknet53 stage. CSPDarknet53 has five stages, and the number of repeated residual modules n is 1, 2, 8, 8, and 4. (b) Structure of an ECA-CSPDarknet44 stage. ECA-CSPDarknet44 has four stages, and the number of repeated residual modules n is 1, 2, 8, and 8.

transition layer will be concatenated and then undergo another transition layer.

There are two problems when directly using CSPDarknet53 as a feature extraction network for tubular solder joint detection. First, in our task, tubular solder joints are very small, and CSPDarknet53 with 32 times downsampling will lead to model optimization difficulties because too large downsampling multiple is easy to cause excessive weakening of object features. Second, CSPDarknet53 lacks adaptive selection ability for important features of the target so that the network consumes on some nonimportant features, which reduces the efficiency of the model.

In order to solve the above two problems, we propose a new feature extraction network named ECA-CSPDarknet44 for tubular solder joint detection. In ECA-CSPDarknet44, we choose Darknet44 with 16 times downsampling, which reduces the downsampling multiples by discarding the last stage in Darknet53. It can effectively reduce and even avoid model optimization difficulties caused by the excessive weakening of object features. In addition, we introduce ECA [39] module into our feature extraction network. It is an effective and advanced channel attention method that could capture local cross-channel interaction and adaptively select important channel features with little increasing of model complexity. As shown in Fig. 3(b), the ECA module could be embedded into the residual module. The process can be described as the following formula:

$$W^* = W \otimes \mathrm{Sigmoid}(\mathrm{Conv1}\,d(\mathrm{Gav}(W))) \tag{1}$$

where Gav is global average pooling, which is used to aggregated channel features. ECA generates channel weights by performing a fast 1-D convolution (Convld). Sigmoid is the activation function. $\otimes$ is the elementwise product. W is the input feature map, and W^* is the output feature map after the ECA module.

B. Cross-Spatial Feature Fusion Network

As shown in Fig. 4, CFFN contains multiscale global feature fusion from different convolutional layers and multiscale local region feature fusion on the same convolutional layer.

Multiscale global feature fusion is widely used in object detection, such as FCN [41], U-Net [42], and DeepLab v3+ [43]. It is verified to improve the conventional detector's performance a lot. In a CNN-based detector, different levels of features in CNN are designed to encode different levels of information. High-level features pay more attention to the semantic information of objects, while lower level features contain more detailed information. The global feature fusion from different stages makes full use of the complementary advantages of different levels to enhance the expression ability of feature maps. The detail of multiscale global feature fusion is shown in Fig. 4; X1 and X2 represent the global feature maps that come from different stages of feature extraction networks, which have different resolutions. A high-resolution feature map goes through a series of transition layers and downsampling processes and then concats with a low-resolution feature map branch on the channel.

In multiscale local region feature fusion, we introduce the improved SPP [44] to integrate multiscale local region features on the same convolutional layer, improving the ability of feature expression further, different from the classical spatial pyramid pooling [45], which divides the input feature map into a fixed-length representation regardless of image scale. The improved SPP uses the same network structure as SPP but is mainly for feature fusion by adjusting the parameters on the pooling layer of SPP. It uses maxpooling with stride = 1 in the spatial pyramid pooling layer and pools the feature maps by the different sliding windows of which the sizes are 5 × 5, 9 × 9, and 13 × 13 separately. Then, three feature maps pooled with size of $\mathrm{size}_{\mathrm{fmap}} \times \mathrm{size}_{\mathrm{fmap}} \times 512$ are concatenated

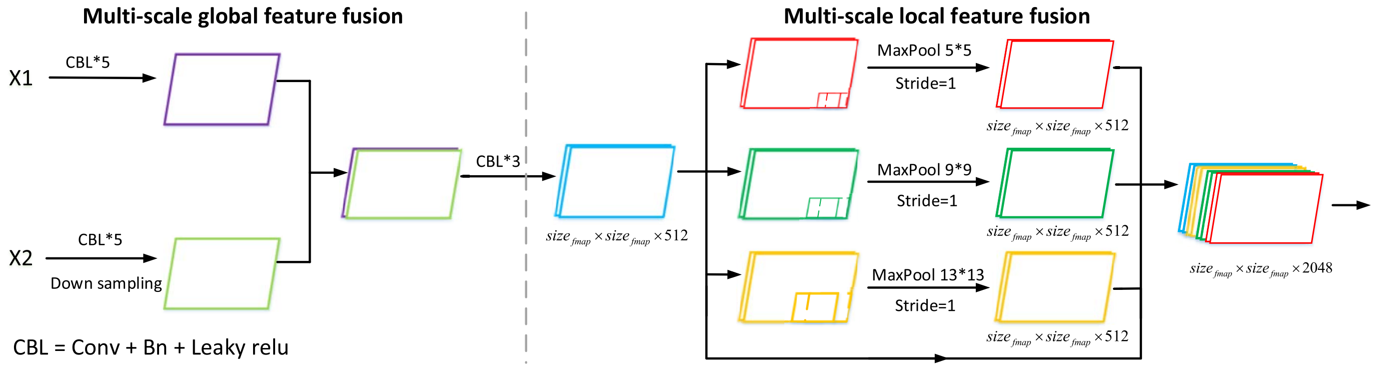

Multi-scale global feature fusion
X1
CBL*5
X2
CBL*5
Down sampling
CBL*3
$size_{fmap} \times size_{fmap} \times 512$
CBL = Conv + Bn + Leaky relu
Multi-scale local feature fusion
MaxPool 5*5
Stride=1
MaxPool 9*9
Stride=1
MaxPool 13*13
Stride=1
$size_{fmap} \times size_{fmap} \times 512$
$size_{fmap} \times size_{fmap} \times 512$
$size_{fmap} \times size_{fmap} \times 512$
$size_{fmap} \times size_{fmap} \times 2048$

Fig. 4. CFFN. It consists of multiscale global feature fusion from different convolutional layers and multiscale local region feature fusion on the same convolutional layer.

with the input feature maps of the SPP block on the channel, producing $\text{size}_{\text{fmap}} \times \text{size}_{\text{fmap}} \times 2048$ feature maps as the output for object detection.

The global feature fusion and the local feature fusion are connected serially in CFFN, greatly enriching the expression ability of feature maps. Multiscale global feature fusion and multiscale local feature fusion can be expressed mathematically by formulas (2) and (3) separately

$$f_1 = \text{Concat}\{\varphi_1(X_1), \varphi_2(X_2), \ldots, \varphi_i(X_i)\} \tag{2}$$

where X_i means the feature maps of different stages in the feature extraction network. $\varphi_i(\cdot)$ means the transformation function of each source feature map before being concatenated together. Concat means feature maps concatenated along the channel. f_1 represents the output feature map after multiscale global feature fusion

$$f_2 = \text{Concat}\left\{\phi_1(f_1), \phi_2(f_1), \ldots, \phi_j(f_1)\right\} \tag{3}$$

where ϕ_1 means the maxpooling process in spatial pyramid pooling layer. Concat is the same with formula (2). f_2 represents the output feature map after multiscale local feature fusion.

C. Bin-Based Rotation Detection Network

We propose a BRDN for tubular solder joints. In current oriented object detection, oriented anchor or dedicated rotating RPN is mostly adopted. These ways need massive computation and greatly increase the complexity of the model. Different from oriented anchor or dedicated rotating RPN, we decompose the prediction of rotation box into horizontal box positioning and accurate angle regression. As our method is based on horizontal anchor, it not only reduces the amount of computation but also easily migrates to other object detection models. In addition, without complex structure, it is realized by coding and decoding, reducing the model complexity effectively.

The label of a tubular solder joint is $(B_x, B_y, B_w, B_h, B_\theta)$. (B_x, B_y), B_w, and B_h represent center point coordinate, width, and height of horizontal box positioning, respectively. For angle B_θ regression, we introduce a bin-based angle regression method [46] from 3-D detection field. It transforms a regression task into a classification and regression task. B_θ is encoded as the bin classification target $\text{bin}_\theta : [\text{bin}_\theta^1, \text{bin}_\theta^2, \ldots, \text{bin}_\theta^m]$ and the residual regression target $\text{res}_\theta : [\text{res}_\theta^1, \text{res}_\theta^2, \ldots, \text{res}_\theta^m]$. The angle of individual accurate regression adjusts the network-predicted horizontal bounding box, and then, the oriented bounding box can be got. Therefore, the loss function of BRDN includes horizontal box positioning, classification loss, confidence loss, and angle regression loss of tubular solder joint. It can be expressed as follows:

$$\text{loss}_{\text{obb}} = \text{loss}_{\text{hbb}} + \text{loss}_{\text{ang}} + \text{loss}_{\text{class}} + \text{loss}_{\text{conf}}. \tag{4}$$

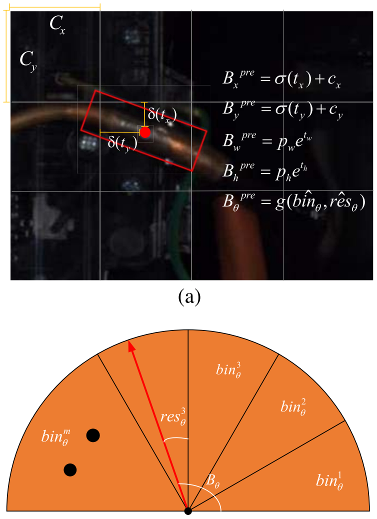
loss_{hbb} and loss_{ang} are used to obtain the target oriented bounding box. $loss_{conf}$ is the confidence loss of the candidate target, which is used to predict whether there is an object in oriented bounding box. $\text{loss}_{\text{class}}$ is the classification loss of the candidate target, which is used to predict the category of the candidate target. In our model, $\text{loss}_{\text{class}}$ uses the softmax cross-entropy, and $\text{loss}_{\text{conf}}$ uses the binary cross-entropy.

We adopt the way of YOLO series [5], [20] for horizontal box positioning. The input image is divided into $S \times S$ grids, and each grid is responsible for the detection of an object. BRDN predicts four coordinates for each horizontal anchor, which is a preset box in each grid, t_x, t_y, t_w, and t_h. (t_x, t_y) and (t_w, t_h) represent the network-predicted center offset of the anchor and the scaling of width and height, respectively. Suppose that the grid is offset from the top left corner of the image by (c_x, c_y), and the anchor prior has width and height p_w and p_h. The conversion from initial anchor to predicted box can be described as $B_x^{\text{pre}} = \sigma(t_x) + c_x$, $B_y^{\text{pre}} = \sigma\left(t_y\right) + c_y$, $B_w^{\text{pre}} = p_w e^{t_w}$, and $B_h^{\text{pre}} = p_h e^{t_h}$. $\sigma(\cdot)$ is the activation function. During model training, loss_{hbb} uses DIoU [47] loss. The mathematical expression is given as follows:

$$\text{loss}_{\text{hbb}} = 1 - \frac{\left|B \cap B^{gt}\right|}{\left|B \cup B^{gt}\right|} + \frac{\rho^2\left(b, b^{gt}\right)}{d^2} \tag{5}$$

where $B^{gt} = \left(x^{gt}, y^{gt}, w^{gt}, h^{gt}\right)$ is the ground-truth, and $B = (x, y, w, h)$ is the predicted box. b and b^{gt} denote the central points of B and B^{gt}, $\rho(\cdot)$ is the Euclidean distance, and d is the diagonal length of the smallest enclosing box covering the two boxes.

In angle regression of tubular solder joint detection, the orientation is defined as solder joint along the right side of metal pipe, and the angle range is $[0, \pi]$. As shown in Fig. 5(b),


C_x
C_y
$\delta(t_x)$
$\delta(t_y)$
$B_x^{pre} = \sigma(t_x) + c_x$
$B_y^{pre} = \sigma(t_y) + c_y$
$B_w^{pre} = p_w e^{t_w}$
$B_h^{pre} = p_h e^{t_h}$
$B_\theta^{pre} = g(\hat{bin}_\theta, \hat{res}_\theta)$
bin_θ^3
bin_θ^2
bin_θ^1
bin_θ^m
res_θ^3
B_θ

(a)

(b)

Fig. 5. Prediction process in BRDN. (b) B_θ is encoded as the bin classification target $\text{bin}_\theta : [\text{bin}_\theta^1, \text{bin}_\theta^2, \ldots, \text{bin}_\theta^m]$ and the residual regression target $\text{res}_\theta : [\text{res}_\theta^1, \text{res}_\theta^2, \ldots, \text{res}_\theta^m]$ in bin-based angle regression. (a) Oriented bounding box prediction process. (b) Bin-based angle regression of tubular solder joint.

we divide angle π into m bins and calculate the bin classification target $\text{bin}_\theta : [\text{bin}_\theta^1, \text{bin}_\theta^2, \ldots, \text{bin}_\theta^m]$ and the residual regression target $\text{res}_\theta : [\text{res}_\theta^1, \text{res}_\theta^2, \ldots, \text{res}_\theta^m]$. Each residual regression head $\text{res}_\theta^j (1 \le j \le m)$ is only responsible for bin_θ^j. m is a hyperparameter. The encoding process of angle θ is shown as follows:

$$\text{bin}_\theta^j = \begin{cases} 1 & , j = \lceil \theta/m \rceil \\ 0 & , j \neq \lceil \theta/m \rceil \end{cases} \tag{6}$$

$$\text{res}_\theta^j = \begin{cases} \theta \% m & , j = \lceil \theta/m \rceil \\ 0 & , j \neq \lceil \theta/m \rceil. \end{cases} \tag{7}$$

In the loss function, the bin classification target uses the cross-entropy cost function; the residual regression target uses the mean squared error cost function. The mathematical expression of loss_{ang} is given as follows:

$$\text{loss}_{\text{ang}} = \frac{1}{N} \sum_i^N \sum_j^m \Big[f\left(\text{res}_\theta^{ij}, \hat{\text{res}}_\theta^{ij}\right) - \text{bin}_\theta^{ij} \log\left(\hat{\text{bin}}_\theta^{ij}\right) \Big] \tag{8}$$

$$f\left(\text{res}_\theta^{ij}, \hat{\text{res}}_\theta^{ij}\right) = \begin{cases} 0, & \text{bin}_\theta^{ij} = 0 \\ \left(\text{res}_\theta^{ij} - \hat{\text{res}}_\theta^{ij}\right)^2, & \text{bin}_\theta^{ij} = 1 \end{cases} \tag{9}$$

where $\hat{\text{bin}}_\theta^{ij}$ and $\hat{\text{res}}_\theta^{ij}$ are the predicted bin assignments and residuals of angle, bin_θ^{ij} and res_θ^{ij} are the encoded ground-truth targets, and N is the number of samples in a batch.

The predicted angle B_θ^{pre} can be decoded from the predicted bin classification target $\hat{\text{bin}}_\theta$ and the predicted residual regression target $\hat{\text{res}}_\theta$ by the following formula:

$$B_\theta^{\text{pre}} = g(\hat{\text{bin}}_\theta, \hat{\text{res}}_\theta) = (j-1) * (\theta/m) + \hat{\text{res}}_\theta^j \quad \{j \mid \forall x : \hat{\text{res}}_\theta^x \le \hat{\text{res}}_\theta^j; x, \quad j \in [1, 2, \ldots, m]\}. \tag{10}$$

D. New Evaluation Index

In object detection, Intersection over Union (IoU) is mainly used to analyze the performance of bounding box regression. That is, when meeting formula (11), the target in the sample will be determined as true positive (TP). IoU_{thr} in formula (11) mostly is set as 0.6. However, IoU reflects the coincidence rate between target box and prediction box. It mainly focuses on the measurement of object positioning, lacking of accurate measurement and analysis for angle regression in tubular solder joint detection.

In order to measure the angle regression of objects more precisely, we add an angle condition to the original IoU condition when measuring the detection results. That is, when both formulas (11) and (12) are met, the target in the sample will be determined as TP. In this way, false positive (FP) and false negative (FN) can also be defined

$$\frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} > \text{IoU}_{\text{thr}} \tag{11}$$

$$\frac{|B_\theta^{\text{pre}} - B_\theta|}{180^\circ} \le \delta_\theta \tag{12}$$

where IoU_{thr} is the threshold in IoU condition. δ_θ is the relative deviation threshold in angle condition. B^{gt} and B_θ are the ground truths. B is the predicted box, and B_θ^{pre} is the predicted angle.

In object detection, main evaluation indices, precision (P) and recall (R), can be calculated by TP, FP, and FN. Recall (R) and precision (P) can be utilized to get F1 and average precision (AP) values of each category. The mAP is calculated by the mean value of AP over all categories. They are defined as follows:

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

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

$$\text{AP} = \int_0^1 \text{P(R)} d\text{R} \tag{15}$$

$$\text{mAP} = \frac{1}{N_{\text{cls}}} \sum_{i=1}^{N_{\text{cls}}} \text{AP}_i \tag{16}$$

$$\text{F1} = \frac{2 * \text{P} * \text{R}}{\text{P} + \text{R}} \tag{17}$$

where P(R) indicates the P-R function [48] and N_{cls} represents the number of categories. Tubular solder joint detection belongs to single-target detection with one target category. Therefore, $N_{\text{cls}} = 1$. In general, P and R are too one-sided for the same detection algorithm. Because of good balance between precision and recall, mAP is widely used to evaluate detection algorithms.

IV. Experiment

A. Experimental Setup

The dataset comes from the industrial production line, which contains 967 images from 967 refrigerators, including 2872 tubular solder joints. Among the dataset, 677 images are

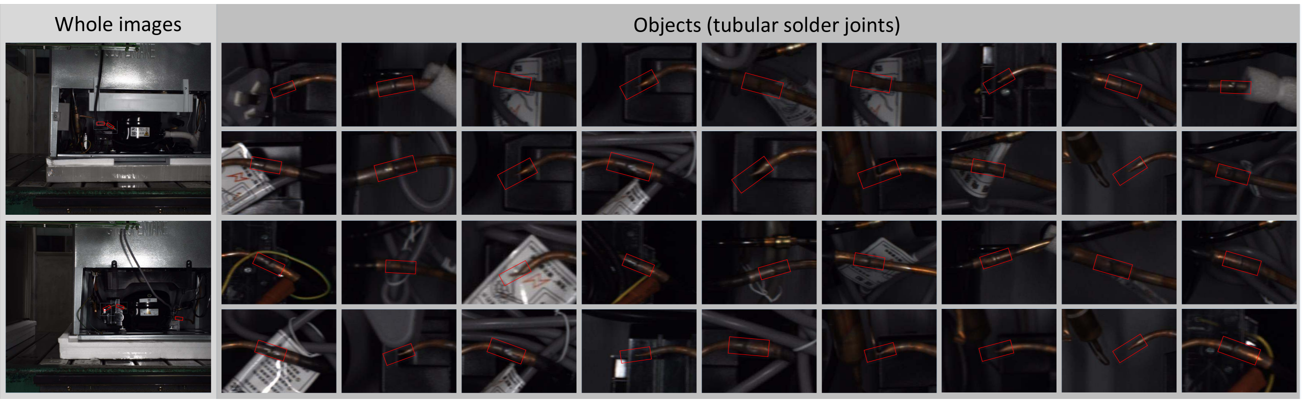


Fig. 6. Examples of the tubular solder joint dataset. The dataset contains 967 images, including 2872 tubular solder joints. Red boxes in the image represent tubular solder joints. The left part shows the whole image, whose resolutions are 2448 × 2048. The right part shows the details of different tubular solder joints. Objects are small and illegible. Meanwhile, the background is complex and low contrast.

randomly selected for training and 290 images for testing. The resolution of each image is 2448 × 2048, and a tubular solder joint is about 750 pixels. The dataset is shown in Fig. 6.

Our designed ACR-Net is an end-to-end learning network. All experiments are trained and tested on an NVIDIA GeForce GTX 1080 Ti GPU with 11-GB memory. As the original picture with 2448 × 2048 resolution is too large, proper image size compression will help model calculation and storage. The original picture is resized by the ratio of 2448 to 608. Then, the short side of the resized image is padded to 608 with a fixed pixel value. Therefore, the original picture is resized from 2448 × 2048 to 608 × 608 without changing the geometric shape of the target and the background. In postprocessing, non-maximum suppression (NMS) is used for network-predicted oriented bounding boxes. In NMS, the object confidence threshold and IoU threshold are set as 0.2 and 0.4 separately. The training step is 34 000 (150 epochs). The step decay learning rate scheduling strategy is adopted with an initial learning rate of 0.01 and multiplied with a factor of 0.1 at the 25 000 steps and the 30 000 steps. We adopt the momentum optimizer as a model optimizer where the weight decay is set as 0.0005 and the momentum is set as 0.9. In all experiments, the batch size is set as 3. For improving the object detection training, ACR-Net uses image flip, CutMix, and Mosaic in data augmentation.

Besides Section IV-G3, all experiments follow the way of original evaluation index in general object detection, and IoU is the only TP judgment condition. The IoU threshold for calculating TP is set as 0.6. In Section IV-G3, we solely analyze the influence of our proposed evaluation index on tubular solder joint detection.

B. Hyperparameter m in Bin-Based Rotation Detection Network

In BRDN, bin-based angle regression has an important influence on oriented bounding box regression. In this experiment, we set different values of m to analyze the impact on the angle regression of tubular solder joint. $m = 6$ means using six bin classification heads and six residual regression heads in angle regression based on ACR-Net. $m = 1$ is equivalent to the way of directly regressing, which uses the mean squared error cost function.

TABLE I
USING DIFFERENT m'S FOR ACR-NET TRAINING

m	P(%)	R(%)	mAP(%)	F1(%)
1	97.3	96.4	95.3	96.8
3	97.9	99.3	98.2	98
6	98	99.5	**98.8**	98.8
9	98.4	98.2	97.9	98.3
12	98.3	97.4	96.7	97.9
15	94.9	94.8	92.8	94.9
18	93.1	93.6	91.2	93.3

As we can see from Table I, ACR-Net gets the best performance with 98.8% mAP when $m = 6$. This experiment shows that bin-based angle regression is much better than directly regressing. The larger m is, the more difficult it is for bin classification target, while the smaller m is, the more difficult it is for residual regression target.

C. Comparison With Other Attention Module in Feature Extraction Network

In the feature extraction network, the attention mechanism is good at capturing the internal correlation of features, automatically identifying the importance of the target features, and adaptively guiding the network to learn the important features of solder joints. In this experiment, in order to analyze the influence of different attention modules on the feature extraction of the tubular solder joint, we introduce different attention module including SE-net [49], CBAM [50], and BAM [51] into feature extraction network, replacing the ECA module. The attention mechanism is mainly divided into channel attention mechanism and spatial attention mechanism. SE-net and ECA belong to channel attention. BAM and CBAM combine channel attention and spatial attention.

As we can see from Table II, the last two lines separately represent CSPDarknet44 or CSPDarknet53 without attention mechanism. In general, the added attention module in the feature extraction network improves the network's detection accuracy of tubular solder joints, because the attention mechanism can select the information that is more critical to the current task goal from huge information. Then, invest more

TABLE II

Using Different Attention Modules for ACR-Net Training on the Tubular Solder Joint Dataset. FPSs Were Measured on the Same Machine

Backbone	P(%)	R(%)	mAP(%)	F1(%)	FPS
SE-CSPDarknet44	99.1	98.4	98.2	98.7	25.8
ECA-CSPDarknet44	98	99.5	**98.8**	98.8	31.3
CBAM-CSPDarknet44	95	99.1	97.5	97	14.8
BAM-CSPDarknet44	95.1	98.4	97.3	96.7	16.9
CSPDarknet44	97.1	96.8	95.8	96.9	34.5
CSPDarknet53	96.9	95.7	95.1	96.3	33.2

attention resources to obtain more detailed features of the target that need to be paid attention to while suppressing other useless features, significantly improving the feature extraction ability for tubular solder joints. CSPDarknet44 works better than CSPDarknet53 because the high downsampling ratio in CSPDarknet53 tends to increase the difficulty of learning the features of tubular solder joints.

The experimental result shows that ECA-CSPDarknet44 achieves optimal accuracy and speed in tubular solder joint detection. The attention module, ECA, works better than BAM, SE, and CBAM. ECA improves the efficiency of feature extraction obviously with a little decrease in speed. The detection accuracies of SE-net, BAM, and CBAM are both higher than the detection accuracy of the model without attention but, due to increased more model complexity, detection speed becomes much slower. Moreover, ECA and SE-net work better than both BAM and CBAM on the whole. In essence, both channel attention and spatial attention are trained with the model together to obtain a weight distribution, and then, this weight can be used to select the channel or superimpose on each pixel of the feature map. The spatial attention mechanism uses each pixel of the feature map as the basic calculation unit to adjust the weight distribution on the feature map. The channel attention mechanism uses a channel layer as the calculation unit to adjust the weight distribution on the channel. Therefore, the spatial attention mechanism is a more computationally consuming process than the channel attention mechanism, which affects the forward propagation time. Meanwhile, in solder joint detection, the channel-level selection of feature maps is more beneficial to the extraction and identification of characteristic information of tubular solder joint than the spatial level. ECA could capture local cross-channel interaction effectively and adaptively select important channel features with little increasing of model complexity. ECA is the optimal choice for tubular solder joint detection comparing with other attention methods.

D. Ablation for Cross-Spatial Feature Fusion Network

CFFN is designed to improve the ability of feature expression for tubular solder joints in ACR-Net. It contains multiscale global feature fusion and multiscale local feature fusion. In this experiment, we make an ablation study for CFFN. Our baseline is ECA-CSPDarknet44 + BRDN, without feature fusion. In addition, we carry out a comparative experiment for multiscale local feature fusion. In detail, we introduce ASPP [55] to replace multiscale local feature fusion in CFFN. ASPP combines atrous convolution and spatial pyramid pooling.

TABLE III

Ablation for CFFN. Baseline Is ECA-CSPDarknet44 + BRDN, Without Feature Fusion. CFFN Consists of Multiscale Global Feature Fusion (F1) and Multiscale Local Feature Fusion (F2)

Method	P(%)	R(%)	mAP(%)	F1(%)	FPS
ECA-CSPDarknet44 + BRDN (baseline)	94.4	94.7	93.4	94.6	33.7
+ F1	97.9	98.6	98.3	98.2	33.6
+ F2	94.8	97.8	96.2	96.2	32.4
+ CFFN (F1+F2)	98	99.5	**98.8**	98.8	31.3
+ CFFN (F1+ASPP)	98	99.1	98.6	98.5	33.4

TABLE IV

Ablation for Loss Function in BRDN. l_{HBB}, l_{ANG}, l_{CONF}, and l_{CLASS} Represent Loss$_{HBB}$, Loss$_{ANG}$, Loss$_{CONF}$, and Loss$_{CLASS}$, Respectively

l_{hbb}	l_{ang}	l_{conf}	l_{class}	P(%)	R(%)	mAP(%)	F1(%)
✓				0	0	0	0
	✓			0	0	0	0
		✓		0	0	0	0
			✓	0	0	0	0
	✓	✓	✓	95.0	2.2	2.4	4.4
✓		✓	✓	11.0	5.8	1.8	7.5
✓	✓		✓	0	0	0	0
✓	✓	✓		98.0	99.5	98.8	98.8
✓	✓	✓	✓	98.0	99.5	98.8	98.8

As we can see from Table III, baseline gets 93.4% mAP with 33.7 FPS. Adding CFFN, our model significantly improves accuracy with little speed reduction. ASPP replacing multiscale local feature fusion can increase the speed with little precision sacrifice. In tubular solder joint detection, multiscale global feature fusion plays a more important role than multiscale local feature fusion. This experiment effectively proves that the CFFN improves the network's ability to express the characteristics of tubular solder joints.

E. Ablation for Loss Function in Bin-Based Rotation Detection Network

BRDN is proposed for oriented bounding box regression through a decoupling approach. The loss function is composed of loss$_{hbb}$, loss$_{ang}$, loss$_{class}$, and loss$_{conf}$. In this experiment, we make an ablation study for the loss function in BRDN.

As we can see from Table IV, loss$_{hbb}$, loss$_{ang}$, and loss$_{conf}$ are the essential elements for oriented bounding box regression in tubular solder joint detection. loss$_{class}$ is essential for the object detection network of multitarget classification, but it is not for tubular solder joint detection with only one target category. As there is only one target category in tubular solder joint detection, loss$_{conf}$ determines that there is an object in oriented bounding box automatically, which means that there is a tubular solder joint in the box. To avoid losing the generality and standardization of our method, we still retain loss$_{class}$ in the loss function.

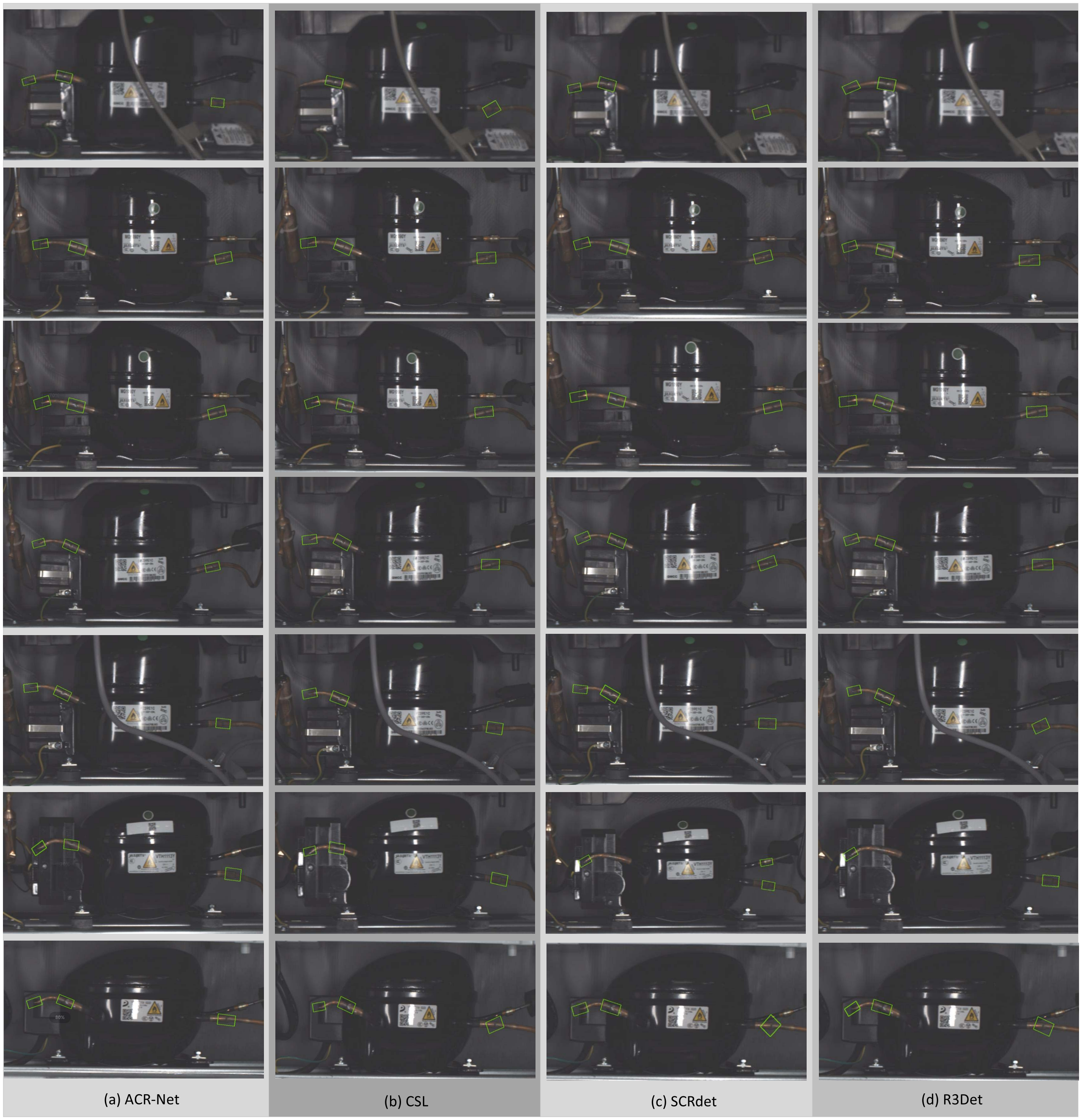


Fig. 7. Visualization of detection results of different oriented object detection methods. (a) Detection results of our proposed ACR-Net. Green boxes in image represent tubular solder joints. (b)–(d) Detection results of major other detection methods, which have these problems of error detection, missed detection, and low detection accuracy. (b) CSL. (c) SCRdet. (d) R3Det.

F. Comparison With State-of-the-Art Oriented Object Detection Methods

To verify the effectiveness of our proposed ACR-Net, we make a comparison with state-of-the-art oriented object detection methods on tubular solder joint detection in this experiment. We compare with R2CNN [52], SCRDet [53], CSL [54], Xu *et al.* [27], and R3Det [28], and these methods are experimented on the same machine, which is the same as ACR-Net. In testing, all the images are resized as 608 × 608.

Table V summarizes the experimental results of different methods. Fig. 7 shows the visualization of detection results of different oriented object detection methods. Our proposed ACR-Net outperforms other advanced rotation methods and is far faster than other methods in detection speed. ACR-Net gets the best accuracy of 98.8% with 31.1FPS on tubular solder joint dataset, meeting industrial application scenario requirements. These general oriented object detectors mainly perform well on some public datasets. However, in specific application scenarios, such as tubular solder joint detection, these detectors do not work well because of having a redundant and unreasonable structural design for tubular solder joint detection. Due to powerful feature extraction ability and feature expression or tubular solder joints, our proposed ACR-Net shows better precision performance. Meanwhile, BRDN in ACR-Net uses

TABLE V

Comparison With State-of-the-Art Oriented Object Detection Methods With Detection Result (x, y, w, h, θ)

Method	P(%)	R(%)	mAP(%)	F1(%)	FPS
ACR-Net	98	99.5	**98.8**	98.8	**31.3**
R2CNN [52]	88.5	88.1	77.6	88.3	4.58
SCRdet [53]	90.8	90.7	86.8	91.2	4.57
CSL [54]	93.8	97.1	89.4	96.7	5.13
Xu et al. [27]	90.5	91.7	86.5	91.1	4.0
R3Det [28]	91.2	90.5	85.3	90.9	4.72

TABLE VI

Comparison With State-of-the-Art Horizontal Object Detection Methods With Detection Result (x, y, w, h). The Proposed ACR-Net Is Without Angle Estimation in This Experiment

Method	P(%)	R(%)	mAP(%)	F1(%)	FPS
ACR-Net (without angle estimation)	98.4	99.3	**98.9**	98.8	31.3
YOLOv3 [20]	98.3	96.4	95.9	97.3	40.6
YOLOv4 [5]	97.0	98.1	97.2	95.2	34.7
CenterNet [56]	98.2	94.2	93.7	95.1	3.7

a decoupling approach for oriented bounding box regression without complex structure, reducing the model complexity to increase detection speed effectively. The experimental results verified the effectiveness of our proposed method.

G. Discussion

In order to further analyze the performance of our proposed method, we design the following four aspects of the discussion.

1) Comparison With State-of-the-Art Horizontal Object Detection Methods: Our proposed ACR-Net is designed with rotated detection head (x, y, w, h, θ) and achieves the state-of-the-art detection performance on tubular solder joint detection. To better verify the comprehensive performance of our proposed network structure, we conduct the experiment to compare with some state-of-the-art horizontal object detection methods on tubular solder joint detection, which contain traditional horizontal detection head (x, y, w, h). We compare with YOLOv3 [20], YOLOv4 [5], and CenterNet [56], and these methods are experimented on the same machine, which is the same as ACR-Net. In testing, all the images are resized as 608 × 608.

Table VI summarizes the experimental results of different methods. Our proposed ACR-Net has competitive performance with horizontal object detection methods on tubular solder joint dataset. It gets the best accuracy while meeting real-time requirements. The experimental results show that the proposed network structure has strong generalization.

2) Experimental Analysis for Different IoU Thresholds: In object detection, IoU is mainly used to analyze the performance of bounding box regression. In the case of the same mAP, the larger IOU threshold is, the better the performance of the algorithm is. IoU threshold is mostly set as 0.5 in some well-known public dataset, such as PASCAL VOC [57] and HRSC2016 [58]. In previous experiments, we set a more strict IoU threshold of 0.6. In this experiment, we set different ranges of IoU threshold to analyze the performance of our proposed method.

TABLE VII

Performances of ACR-Net in Different IoU Thresholds

IoU_{thr}	P(%)	R(%)	mAP(%)	F1(%)
0.5	98	99.5	98.8	98.8
0.6	98	99.5	98.8	98.8
0.7	98	99.5	98.8	98.8
0.8	98	99.5	98.8	98.8
0.85	98	99.5	98.8	98.8
0.875	97.8	99.3	98.8	98.5
0.9	93.3	94.7	91.1	94.0
0.925	80.9	82.2	69.9	81.5
0.95	46.4	47.1	26.9	46.7
0.975	0.925	0.939	0.0536	0.932

TABLE VIII

mAP With Different Ranges of $\mathrm{IoU_{thr}}$ and δ_θ in New Evaluation Index. Detection Model Is Our Proposed ACR-Net

mAP(%)		IoU_{thr}					
		0.2	0.4	0.6	0.8	0.9	0.95
δ_θ	0.01	24.6	24.6	24.6	24.6	24.6	19.9
	0.03	86.1	86.1	86.1	86.1	86.1	26.9
	0.05	98.8	98.8	98.8	**98.8**	91.1	26.9
	0.1	99.8	99.8	99.3	99.3	91.1	26.9

As we can see from Table VII, the precision begins to decrease after $\mathrm{IoU_{thr}} = 0.85$. It further illustrates the high accuracy and robustness of our detection algorithm.

3) Experimental Analysis for New Evaluation Index: In all previous experiments, IoU was used as the only condition for judging TP because our proposed ACR-Net is accurate enough in tubular solder joint detection, resulting in that the traditional calculation method of mAP is almost equivalent to our new evaluation index when meeting industrial precision. In this experiment, we set different ranges of $\mathrm{IoU_{thr}}$ and δ_θ to analyze its influences on the results of tubular solder joint detection.

As we can see from Table VIII, in each line, mAP of ACR-Net obviously declines after $\mathrm{IoU_{thr}} = 0.8$, which indicates that the detection results of our proposed method have high coincidence rate with ground truth of targets. It is further verified that our method is robust and effective. In each column, δ_θ is smaller, the more strict it is for the angle regression. δ_θ is adjustable for different precision requirements from different applications. In tubular solder joint detection of the industrial operating system, $\delta_\theta = 0.05$ has been met the demand angle precision requirement basically.

4) Results on Public Dataset: ACR-Net achieves state-of-the-art performance in tubular solder joint detection. In order to explore the possibility of applying ACR-Net to other oriented object detection tasks, we conduct extensive experiments on the HRSC2016 dataset, which contains a large number of thin and long ships with arbitrary orientation. The results show that ACR-Net achieves 84.9 mAP with 14.3 FPS on the HRSC2016 dataset. Comparing with other state-of-the-art methods, such as MFIAR-Net [59], RoI transformer [7], and R3Det [28], ACR-Net achieves comparable accuracy and faster speed on the HRSC2016 dataset. These state-of-the-art methods on the HRSC2016 dataset mainly focus on detection precision and do not particularly concern about detection efficiency. That is caused by the different requirements of different detection tasks. Ship detection of aerial images in the HRSC2016 dataset has no strict real-time requirement, while tubular solder joint detection in industrial systems is required necessarily.

V. Conclusion

In this article, we propose a tubular solder joint detection method named ACR-Net, which could effectively detect solder joint positioning and angle estimation. In addition, we establish a tubular solder joint dataset and conduct sufficient experiments to verify the effectiveness of our method. Our proposed ACR-Net achieves 98.8% mAP with 31.3 FPS on the dataset, meeting the high-precision and real-time requirements of industrial systems. Meanwhile, ACR-Net is far faster and more accurate than other oriented object detection algorithms, achieving state-of-the-art performance on tubular solder joint detection.

References

[1] S. Y. Wang, Y. Zhao, and L. Wen, "PCB welding spot detection with image processing method based on automatic threshold image segmentation algorithm and mathematical morphology," *Circuit World*, vol. 42, no. 3, pp. 97–103, Aug. 2016.

[2] H. Wu and X. Xu, "Solder joint inspection using eigensolder features," *Soldering Surf. Mount Technol.*, vol. 30, no. 4, pp. 227–232, Sep. 2018.

[3] N. Cai, G. Cen, J. Wu, F. Li, H. Wang, and X. Chen, "SMT solder joint inspection via a novel cascaded convolutional neural network," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 8, no. 4, pp. 670–677, Apr. 2018.

[4] Y.-M. Chang, C.-C. Wei, J. Chen, and P. Hsieh, "An implementation of health prediction in SMT solder joint via machine learning," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Apr. 2019, pp. 1–4.

[5] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*. [Online]. Available: http://arxiv.org/abs/2004.10934

[6] H. Wu, W. Gao, and X. Xu, "Solder joint recognition using mask R-CNN method," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 10, no. 3, pp. 525–530, Mar. 2020.

[7] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2849–2858.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[10] D. Xi, Y. Qin, J. Luo, H. Pu, and Z. Wang, "Multipath fusion mask R-CNN with double attention and its application into gear pitting detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[11] S. Tian and K. Xu, "An algorithm for surface defect identification of steel plates based on genetic algorithm and extreme learning machine," *Metals*, vol. 7, no. 8, p. 311, Aug. 2017.

[12] R. Ren, T. Hung, and K. C. Tan, "A generic deep-learning-based approach for automated surface inspection," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 929–940, Mar. 2018.

[13] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1493–1504, Apr. 2020.

[14] Z. Liu, B. Yang, G. Duan, and J. Tan, "Visual defect inspection of metal part surface via deformable convolution and concatenate feature pyramid neural networks," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9681–9694, Dec. 2020.

[15] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[17] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," 2016, *arXiv:1605.06409*. [Online]. Available: http://arxiv.org/abs/1605.06409

[18] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 821–830.

[19] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7263–7271.

[20] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[21] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[23] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850

[24] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 734–750.

[25] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[26] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9657–9666.

[27] Y. Xu *et al.*, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[28] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," 2019, *arXiv:1908.05612*. [Online]. Available: http://arxiv.org/abs/1908.05612

[29] L. Liu, Z. Pan, and B. Lei, "Learning a rotation invariant detector with rotatable bounding box," 2017, *arXiv:1711.09405*. [Online]. Available: http://arxiv.org/abs/1711.09405

[30] Z. Zhang, W. Guo, S. Zhu, and W. Yu, "Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1745–1749, Nov. 2018.

[31] W. Qian, X. Yang, S. Peng, Y. Guo, and J. Yan, "Learning modulated loss for rotated object detection," 2019, *arXiv:1911.08299*. [Online]. Available: http://arxiv.org/abs/1911.08299

[32] S. M. Azimi *et al.*, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 150–165.

[33] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[34] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.

[35] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5909–5918.

[36] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1962–1969.

[37] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: Fast oriented text spotting with a unified network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5676–5685.

[38] H. Wang *et al.*, "All you need is boundary: Toward arbitrary-shaped text spotting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12160–12167.

[39] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

[40] C.-Y. Wang, H.-Y. Mark 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.

[41] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[42] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Cham, Switzerland: Springer, 2015.

[43] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, *Encoder-Decoder With Atrous Separable Convolution for Semantic Image Segmentation*. Cham, Switzerland: Springer, 2018.

[44] Z. Huang, J. Wang, X. Fu, T. Yu, Y. Guo, and R. Wang, "DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection," *Inf. Sci.*, vol. 522, pp. 241–258, Jun. 2020.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[46] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 770–779.

[47] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proc. AAAI*, 2020, pp. 12993–13000.

[48] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[49] 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.

[50] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[51] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*. [Online]. Available: http://arxiv.org/abs/1807.06514

[52] Y. Jiang *et al.*, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*. [Online]. Available: http://arxiv.org/abs/1706.09579

[53] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8232–8241.

[54] X. Yang, J. Yan, and T. He, "On the arbitrary-oriented object detection: Classification based approaches revisited," 2020, *arXiv:2003.05597*. [Online]. Available: http://arxiv.org/abs/2003.05597

[55] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[56] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[57] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2009, doi: 10.1007/s11263-009-0275-4.

[58] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1074–1078, Aug. 2016.

[59] F. Yang, W. Li, H. Hu, W. Li, and P. Wang, "Multi-scale feature integrated attention-based rotation network for object detection in VHR aerial images," *Sensors*, vol. 20, no. 6, p. 1686, Mar. 2020.

Chenlin Zhou received the B.S. degree from the School of Mechatronics and Vehicle Engineering, Chongqing Jiaotong University, Chongqing, China, in 2019. He is currently pursuing the M.Sc. degree with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.

His current research interests include robot dexterous grasping, computer vision, and deep learning.

Daheng Li received the B.S. degree from the Tongji Medical College, Huazhong University of Science and Technology, Hubei, China, in 2020. He is currently pursuing the M.Sc. degree with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.

His current research interests include reinforcement learning, computer vision, and robot control.

Peng Wang (Member, IEEE) received the B.Sc. degree in electrical engineering and automation from Harbin Engineering University, Harbin, China, in 2004, the M.Sc. degree in control science and engineering from the Harbin Institute of Technology, Harbin, in 2007, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010.

In 2010, he joined the Institute of Automation, Chinese Academy of Sciences, where he is currently a Professor. He has published more than 80 journal articles and conference papers. His current research interests include robotic vision, robotic learning, robotic manipulation, neurorobotics, brain-like robotics, and intelligent robot systems.

Dr. Wang was awarded the First Prize of Science and Technology of Beijing, the Second Prize of Science and Technology of Beijing, and the "Outstanding Reviewers of 2020" of IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT.

Jia Sun received the B.Sc. degree in the measurement and control technology and instrument from the North University of China, Taiyuan, China, in 2009, the M.Sc. degree in instrument science and technology from the Beijing Institute of Technology, Beijing, China, in 2012, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2018.

She is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. Her research interests include robotic vision, vision inspection, and intelligent robotics.

Yikun Huang received the B.S. degree from the School of Automation, University of Science and Technology Beijing, Beijing, China, in 2017, and the M.Sc. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2020.

His current research interests include computer vision, industrial inspection, and medical image diagnosis.

Wanyi Li received the M.Eng. degree in computer science from Guizhou University, Guiyang, China, in 2010, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014.

He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include computer vision and robot learning.