Surface defects inspection of cylindrical metal workpieces based on weakly supervised learning

Mu Ye · Weiwei Zhang · Guohua Cui · Xiaolan Wang

Received: 6 June 2021 / Accepted: 17 November 2021 / Published online: 2 December 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract
Weakly supervised learning applies image tag labels to train convolutional neural networks to locate defect. In industrial vision system, metal surface is anisotropic under light in all directions and it is inevitable to cause local overexposure due to the natural reflection of active strong light, especially on the cylindrical metal surface. In this paper, injector valve is taken as the representative of cylindrical metal workpieces. Since the variety and complexity of cylindrical metal workpiece defects which cause pixel-level annotation require expensive manual work. This problem hinders the application of convolutional neural network in industries. In order to solve these above challenges, this paper proposed an end-to-end weakly supervised learning framework named Integrated Residual Attention Convolutional Neural Network (IRA-CNN). IRA-CNN only uses image tag annotation for training and performs defect classification and defect segmentation simultaneously. Weakly supervised learning is achieved by extracting category-related spatial features from defect classification scores. IRA-CNN is composed of multiple Integrated Residual Attention Block (IRA-Block) as the feature extractor which improves the accuracy and achieves real-time performance. IRA-Block adds Integrated Attention Module (IAM) which includes channel attention submodule and spatial attention submodule. The channel attention submodule adaptively extracts the channel attention feature map to improve its bilateral nonlinearity and the robustness. IAM can be well integrated into the IRA-CNN makes the neural network suppress the interference of useless background area and highlight the defect area. Satisfied performance is achieved by the proposed method in our own defect dataset which could meet the requirements in the industrial process. Experimental results show that the method has good generalization ability. The accuracy of defect classification reaches 97.84% and the segmentation accuracy is significantly improved compared with the benchmark method.

Keywords
Machine vision · Defect detection · Deep learning · Weakly supervised learning · Convolutional neural network

1 Introduction
Cylindrical metal workpiece needs to be matched with other parts in kinematic pairs. Its surface adhesion directly affects the performance of workpiece and even mechanical system. The injector valve, as the representative of cylindrical metal workpieces, has multiple surface quality hazards in its total lifecycle. The surface quality of the injector valve affects the open/close dynamic performance of Liquid pressure circuit. The surface defects of injector valve will have a serious negative impact on the engine injector and reduce automotive service life. Therefore, in the manufacturing process of injector valve, defect inspection is vital for production. Due to the improper manufacturing process and raw material quality, there are some defects on the surface of injector valve, such as surface Dirt, Scratch, and Electrochemical Corrosion (EC). In recent years, machine vision has gradually replaced the manual detection and been applied to automatic defect inspection, which can improve the quality and life of products by automatically identifying product defects [1–3].

The surface image of injector valve acquired by image acquisition system is shown in Fig. 1. Section 3.1 will introduce this system. This paper is dedicated to detecting three types of defects with different shape features: Dirt, Scratch,
EC, and segment dirt area and scratch area. Compared with the normal injector valve, the dirt area of the valve has smaller gray value and irregular shape. The main characteristics of the scratch defects are that the gray value is similar to the normal area, the contrast ratio is low, and the defects are distributed in strip. The defect of EC is characterized by small pits in large area of the injector valve, and according to the experimental observation, EC will cause pits in all areas of the whole injector valve, segmentation is meaningless for the defect of EC. Therefore, EC is not necessary to be processed by image segmentation. These defects have a negative impact on the quality of the injector. The above defects are difficult to be found by manual observation, and some surface defects can only be observed by high resolution camera. Therefore, it is necessary to detect the surface defects of injector valve for product quality control.

The existing surface defect inspection methods based on machine vision mainly targeted at four types of surface: (1) non-textured surface; (2) repeated pattern surface; (3) uniform textured surface; (4) non-uniform textured surface. Cylindrical metal workpieces can be classified as non-textured surface. For detecting this kind of surface defects, the traditional machine vision methods generally include two stages of feature extraction and pattern classification. The features that can be used for pattern classification include statistical measures such as first-order moments and second-order moments, as well as handcrafted features such as Local Binary Patterns (LBP) [4], Gray level Co-occurrence Matrix (GLCM) [5]. Feature classifier includes generating model such as Bayesian classifier and discriminant model such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM) [6], Random Forest [7]. The performance of defect inspection relies to some extent on how well the features are designed. Feature extraction in traditional methods mainly relies on manually designed descriptors, which requires professional knowledge and a complicated parameter adjustment process. Moreover, each method target a specific industrial scenario and has poor robustness and generalization ability.

In recent years, deep learning method has shown its advantages in defect segmentation [8, 9], defect detection and classification [10, 11, 13–16, 18]. Deep learning has the characteristics of high precision and wide application by automatically extracting image features. In the field of
defect inspection, the common surface defect inspection tasks include fabric defect detection, steel defect detection, solar panel defect detection, and LED chips defect detection. In this field, deep learning can be mainly applied to defect classification and detection and defect segmentation. Aiming at the problem of defect classification and detection, Xu [11] proposed roller bearing defect classification based on SDD-CNN. This method uses SSAD method for data expansion and uses InceptionV3 [12] as the classification network, and its classification accuracy reaches 99.56%. Chen et al. [13] proposed a multispectral CNN structure for surface defect classification of solar cells. Three convolutional neural networks are established for three spectral channels for classification, and the classification accuracy is 88.24%. Cheon et al. [14] proposed a CNN structure for wafer surface defect classification, which can detect unknown defects by clustering the eigenvectors of the same defect types. He et al. [15] proposed a semi-supervised defect classification method to make up for the shortage of samples in supervised training. This method uses GAN to generate a large number of unlabeled data. However, the above methods only classify defects and do not get the location information of defects and the feature information of defects (such as size, distribution), so some researchers add object detection and image segmentation methods to defect detection. Li et al. [16] used the improved YOLO [17] network to detect six kinds of steel surface defects, and achieved 97.55% mAP and 95.86% recall. Su et al. [18] designed a novel bidirectional attention feature pyramid network structure and embedded it into the regional proposal network to improve the efficiency of Faster-RCNN [19] in detecting surface defects of solar cell. The target detection method needs to use the label with bounding box for training.

More accurate defect boundary can be obtained by using image segmentation. Tabernik et al. [20] proposed a segmentation framework for crack segmentation, which first segments the defect and then classifies defects combined with the segmentation feature map. The experimental results show that the average accuracy of this method is at least 1.9% higher than that of deeplabv3 + [21] and u-net [22]. Tao et al. [23] used an encoder-decoder structure to segment defects. After that, the segmented region is cropped out for classification. Wang M and Cheng J C P. [24] transformed the inference of CRF into CNN operation, which fully combined CRF with deep convolutional neural network and significantly improved the defect segmentation effect. However, the variety and complexity of cylindrical metal workpieces defects and image contrast fluctuation, making samples annotating time-consuming and be of high cost which caused to hard to use supervised learning.

Based on the above shortcomings, weakly supervised learning can effectively solve the leakage problem of labeled defect samples and realize the detection or segmentation of defect image only by using image-level annotation. Among them, CAM [25] is an important method to realize weakly supervised learning, and CAM can generate a saliency map to show the probability that each pixel belongs to a certain defect category. Lin et al. [26] used CAM method to visually predict the depth CNN and thus locate the LED defect location. Chen et al. [27] proposed a robust weakly supervised learning method for surface defect detection, which uses transfer learning to get CAM. Xu et al. [28] proposed a weakly supervised detection framework, in which CNN model was trained to identify surface cracks in motor commutators. This method achieved 99.5% recognition accuracy in Kolektor SSD dataset. The above researches all use CAM method, but CAM needs to replace the fully connected layer with global average-pooling layer and then retraining, which leads to the high training cost of the model and limits the use scenarios of the model. Grad-CAM [29] and Grad-CAM++ [30] can solve the above problems. The above two methods use the gradient information of the last layer of CNN feature map to assign values to each neuron. The difference lies in that Grad-CAM++ uses a weighted combination of the positive partial derivatives of the last convolutional layer feature maps with respect to a specific class score as weights to generate a visual explanation for the corresponding class label. For the case of multiple defects in one image, Grad-CAM++ can produce more accurate saliency maps. Therefore, Grad-CAM++ has better robustness and interpretability.

The saliency map generated by Grad-CAM++ still has a large noise effect on defect segmentation. Adding attention module in the backbone can effectively alleviate the problem. Attention module can emphasize the weak features of the defect, and suppress the noise in the image. Due to the complex surface and uneven light distribution of the cylindrical metal workpieces, adding attention module is suitable for the surface defect inspection of the injector valve. Attention module includes channel attention module and spatial attention module. These two methods realize “what to look” and “where to look” by calculating the weight of channel direction and spatial position to the feature map. Channel attention module can make CNN network select the feature map related to the defect, so as to suppress useless background information. Spatial attention module can make CNN network pay attention to the defect location information in the feature map. At present, attention mechanism has been widely used in various tasks [31–33]. Hu et al. [31] used global mean
aggregation characteristics to calculate channel direction attention in their squeeze and excitation modules. Ji et al. [32] used attention module to align the context information between the feature maps at different scales and the final prediction of the saliency map. Woo et al. [33] proposed Convolutional Block Attention Module (CBAM), which infers the attention map in turn along two independent dimensions (channel and space), and then multiplies the attention map by the input feature graph to refine the adaptive feature. Inspired by the structure of CBAM, this paper proposes the integrated attention module (IAM). IAM enables classifiers to focus more on defect-related feature maps and regions.

This paper presents a framework for detecting the defects on the outer cylindrical metal surface of fuel injector valve, the contributions of this paper are as follows.

1. Aiming at the defect inspection of cylindrical metal workpieces, an Integrated residual attention convolutional neural network (IRA-CNN) was proposed. IRA-CNN is composed of multiple Integrated Residual Attention Block (IRA-Block). Two residual maps are included in IRA-Block to improve the robustness. Every IRA-Block combines a distinctive attention module to improve the interpretability of network. Therefore, IRA-CNN has a good classification effect for the defects of cylindrical metal workpiece.

2. We proposed IAM which is composed of a channel attention submodule and a spatial attention submodule. In the channel attention submodule, the input proportion of Global Average Pooling (GAP) layer and Global Max-Pooling (GMP) layer was adaptively selected by IAM to adjust the output of different IRA-Block. IAM makes full use of the comprehensive characteristics of GAP layer and GMP layer to improve bilateral non-linearity and robustness of IRA-CNN. Therefore, the background of the image and highlight the defect area of the cylindrical metal workpiece can be suppressed by IAM.

3. In this paper, a weakly supervised learning strategy was proposed to segment the defects of injector valve. This strategy can segment the defects only by using image tag annotation. After classification by IRA-CNN, the saliency map is generated by Grad-CAM++. The pixel-level segmentation of defects is based on the saliency map, which greatly simplifies the task of image segmentation. The time of pixel-level labeling is saved, and the segmentation accuracy is better for the defects of cylindrical metal workpieces.

The rest of this study is as follows. Section 2 introduces the methodology in detail. Section 3 introduces the experiments and related performance evaluation. Finally, the fourth part is the conclusion and discussion.

2 Methodology

In this section, the research of cylindrical metal workpiece defects inspection is mainly based on the injector valve. Surface defects inspection of cylindrical metal workpiece based on weakly supervised learning structure is shown in Fig. 2. We mainly introduce the architectures of the classification module, Integrated attention module, and segmentation module.

2.1 Classification module

Convolutional neural network extracts local feature information from image by convolutional layers and compresses feature information by pooling layers. When detecting the surface defects of injector valve, the sliding window is used to process the original image and take the image divided into eight pieces, which can detect the defects more accurately without reducing the resolution of the original image.

In this paper, IRA-CNN was proposed to extract the features of the outer surface of the injector valve. IRA-CNN is composed of multiple IRA-Blocks in series, in which IAM is integrated. In Section 2.2, the structure of IAM will be explained in detail. EfficientNet [34] and MobileNet [35] also have similar structure, that is, the attention module is integrated into the convolution extraction block. In the IRA-Block, we used $3 \times 3$ convolution kernel to extract features, and stack multiple $3 \times 3$ convolution kernels to expand Receptive field. IRA-Block is shown in Fig. 3.

The function of convolutional layer is to extract the feature of a local region. Different convolution kernels are equivalent to different feature extractors. The defect features of different positions are extracted by sliding the convolutional block in the image. The convolution function is as follows

$$Y = \Phi (W \otimes X + b)$$  \hspace{1cm} (1)

where $\otimes$ denotes element-wise multiplication. $X$ is the input image. $W$ represents the weight of the convolution kernel. $b$ is the bias of the convolution kernel. After convolution operation, the output $Y$ is obtained by using nonlinear activation function $\Phi(\cdot)$.

In order to obtain a larger receptive field and reduce parameter, four $3 \times 3$ convolutional blocks were stacked to extract features. The number of convolution kernel channels is doubled compared with the previous IRA-Block. After feature extraction, IAM is used to suppress the background noise and highlight the defect area. Finally, Max-pooling
is used to down sample and compress features. There is a residual mapping between the input and output of IRA-Block.

IRA-CNN uses four IRA-Blocks to extract the features continuously, the last feature map is sent to the fully connected layer for feature aggregation. The last layer of the fully connected layer is the output of the whole network. The output has the same number of neurons as the labels and can be classified by Softmax classifier. Softmax classifier formula is as follows

$$P(Y_i) = \frac{\exp(Y_i)}{\sum_{i=1}^{K} \exp(Y_i)}$$  \hspace{1cm} (2)

$Y_i$ represents the input of the Softmax classifier while $P(Y_i)$ represents the output probability. $K$ represents the total number of defect categories.

The above is the basic structure of IRA-CNN. IRA-CNN can effectively extract the defect features of injector valve, IRA-CNN network structure is shown in Table 1.

### 2.2 Integrated attention module

Because of the randomness of the surface defect location of the injector valve, IAM is proposed and integrated into the IRA-Block. The proposed IAM cascades the channel attention submodule with spatial attention submodule. The
input and channel attention map contain a residual connection. Relatively, channel attention map and spatial attention map also contain a residual connection. IRA-CNN can suppress the unnecessary background area and highlight the spatial location of the surface defects of the injector valve.

IAM is inspired by CBAM. CBAM uses channel attention modules to cascade spatial attention modules. In channel attention modules, CBAM compresses the spatial information of feature maps through GMP layer and GAP layer, and then sends GAP and GMP into a shared multi-layer perceptron (MLP) with a hidden layer respectively to generate corresponding feature vectors, and then the output feature vectors are merged by using element-wise summation. CBAM compulsorily add the GAP layer and GMP layer after MLP feature extraction cannot well combine the information of GAP layer and GMP layer. The reasons are as follows: CBAM is biased to GMP. Shared layer was used by GAP and GMP to extract features in CBAM, the inference part can be regarded as GAP and GMP added directly and sent to shared layer. Generally speaking, the value of GMP is larger than GAP. The reason is that the ReLU activation function will filter out a large number of negative values, resulting in many values of 0 in the feature map, which leads to the decline of GAP. Therefore, it is biased for CBAM to directly send GAP and GMP into MLP without feature processing. The proposed IAM multiplies GAP layer and GMP layer by trainable weights and then sends them to MLP layer. IAM can make channel attention network adaptively select the proportion of GAP or GMP input to MLP. This channel attention module can get channel weights after only one inference. Moreover, residual connection is added between input and channel attention map, and between channel attention map and spatial attention map, which makes network training easier to converge and improves the bilateral nonlinearity and the robustness of networks. These are the reasons why IAM can achieve better performance than CBAM.

As shown in Fig. 4, given an intermediate feature map $M_{input} \in \mathbb{R}^{C \times H \times W}$ as input and $M_{IAM}$ is the final output attention map. $C$ represents the number of channels in the current feature map. $H$ and $W$ represent the height and width of the current feature map respectively. All operations above can be expressed by the following formula

$$M_{ICAM} = (F_{CW} \otimes M_{input}) + M_{input}$$

$$M_{IAM} = (F_{SW} \otimes M_{CAM}) + M_{ICAM}$$

where $\otimes$ denotes element-wise multiplication. $F_{CW} \in \mathbb{R}^{C \times 1 \times 1}$ and $F_{SW} \in \mathbb{R}^{1 \times H \times W}$ represent channel attention weight and spatial attention weight respectively. $M_{ICAM}$ is channel attention map. IAM makes IRA-CNN have better classification performance and model interpretability, and improves the accuracy to locate defects.

### 2.2.1 Integrated channel attention submodule

In this paper, an integrated channel attention submodule was proposed to obtain the channel attention feature map. This submodule extracts information from squeezing each channel feature map, so that the feature layer with stronger semantic information has higher weight. The ways of squeeze include GAP and GMP, which are proved in [25] and [36] to improve the representation ability of the network. Aiming at the cylindrical metal workpiece inspection, GAP and GMP have different representation ability. GMP focuses on the most significant region in the image to compensate the global region which GAP focus on. Therefore, this paper thinks that GAP and GMP should be jointly input into a network. We multiply GAP and GMP by a trainable weight in the input stage and then add them, and then send them to a network to extract information. Finally, a residual connection is added between the input and output, which is helpful to the convergence of the network.

The specific implementation is as follows, as shown in Fig. 4. Firstly, the input feature map is squeezed into GAP layer $F_{gap} \in \mathbb{R}^{C \times 1 \times 1}$ and GMP layer $F_{gmp} \in \mathbb{R}^{C \times 1 \times 1}$. Set a parameter $\alpha$ and set the weight of GAP to $\frac{e^\alpha}{\alpha+1}$. $\frac{1}{\alpha+1}$ is greater than 0 and no more than 1. Relatively, $\frac{1}{\alpha+1}$ is the weight of $F_{gmp}$. $e$ is the natural constant. So we can ensure that the weight of training is positive. Then, $\frac{e^\alpha}{\alpha+1}F_{gmp}$ and $\frac{1}{\alpha+1}F_{gap}$ are added and sent to MLP to extract information. The hidden layer of MLP contains $C/k$ neurons, $C$ is the number of channels after merging the GAP layer and GMP layer and $k$ is the reduction rate. Finally, sigmoid function is used to activate the last layer to get the channel weight $F_{CW}$. The integrated channel attention map $M_{ICAM}$ is obtained by
multiplying the weight with $M_{\text{input}}$ by element. The final channel attention map also contains a residual connection. The calculation of $M_{\text{ICAM}}$ can be simplified as the following formula

$$FCW = \text{sigmoid} \left( MLP \left( \frac{1}{\varepsilon + F_{\text{gap}}} \right) \right)$$

(5)

$$M_{\text{ICAM}}' = FCW \otimes M_{\text{input}}$$

(6)

$$M_{\text{ICAM}} = M_{\text{ICAM}}' + M_{\text{input}}$$

(7)

where $MLP$ represents that send a set of feature vector into a shared MLP respectively to generate corresponding feature vector.

### 2.2.2 Spatial attention submodule

Spatial attention submodule can make the network focus on “where” the defect is, which is a supplement to channel attention submodule. The input of the spatial attention submodule is $M_{\text{ICAM}}$. Firstly, the average-pooling and max-pooling along the channel axis are applied to calculate two feature maps $F_{\text{avg}}$ and $F_{\text{max}}$; then, the two feature maps are connected and input into a convolution layer to extract features and activate them with activation function. The convolution kernel size is $7 \times 7$, and the output of the convolution layer is spatial weights $F_{\text{SW}}$. $F_{\text{SW}}$ and $M_{\text{ICAM}}$ are multiplied by the corresponding elements. Before getting the final output $M_{\text{IAM}}$, there is a residual connection between $M_{\text{IAM}}$ and $M_{\text{ICAM}}$. The above operation can be simplified as the following formula

$$F_{\text{avg}} = \text{Avgpool} \left( M_{\text{ICAM}} \right)$$

(8)

$$F_{\text{max}} = \text{Maxpool} \left( M_{\text{ICAM}} \right)$$

(9)

$$F_{\text{SW}} = \text{sigmoid} \left( f^{7\times7} \left( [F_{\text{avg}}; F_{\text{max}}] \right) \right)$$

(10)

$$M_{\text{IAM}} = F_{\text{SW}} \otimes M_{\text{ICAM}} + M_{\text{ICAM}}$$

(11)

where $\text{Avgpool}$ and $\text{Maxpool}$ are average-pooling operation and max-pooling operation along the channel axis respectively. $f^{7\times7} \cdot (*)$ represents the convolution of the feature map using $7 \times 7$ convolution kernel.

### 2.3 Segmentation module

#### 2.3.1 Saliency map by grad-CAM++

IAM has made classification network learn “where to look” and “what to look”. Therefore, saliency map is generated by Grad-CAM++ based on the current defect classification score. The probability of whether a pixel value belongs to a defect can be ensured in saliency map. Therefore, saliency map is not only the basis to judge whether the network really recognizes the defect feature, but also an important basis for segmentation.

Grad-CAM++ uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron for a particular decision of interest. The contribution of each pixel value in the feature map to the classification score is obtained by gradient conduction. The method of obtaining gradient weight is different from that of Grad-CAM. Detailed derivation of gradient weight in [29]. Grad-CAM fails to properly localize defects in an image if the image contains multiple occurrences of the same class. Grad-CAM++ solves this problem. Review the calculation method of Grad-CAM++. Suppose that the classification score of a defect classification is $S_c$ (before Softmax operation), $(i, j)$ and $(a, b)$ are iterators over the same activation map $A^k$ are used to avoid confusion. The gradient weight $w^c_k$ of the feature map is calculated as follows

$$\alpha^c_{ij} = \frac{\frac{\partial^2 S_c}{\partial A^k_{ij}}}{\frac{\partial^2 S_c}{\partial A^k_{ij}} + \sum_a \sum_b A_{ab} \left( \frac{\partial^3 S_c}{\partial A^k_{ij}} \right)^2}$$

(12)

$$w^c_k = \sum_i \sum_j \alpha^c_{ij} \text{relu} \left( \frac{\partial S_c}{\partial A^k_{ij}} \right)$$

(13)
Finally, the gradient weight of the feature map is multiplied by the feature map to obtain the saliency map. The formula is as follow

$$L_{ij}^c = \sum_k w_k^c \cdot A_{ij}^k$$  \hspace{1cm} (14)$$

where $L_{ij}^c$ represents saliency map, which can visualize the importance of each pixel value to the current classification score. Therefore, saliency map can be used for defect segmentation.

### 2.3.2 Segmentation framework

The saliency map obtained by Grad-CAM++ represents the probability of whether each pixel is a defect, so it is necessary to binarize the saliency map to obtain the contour of the defect area. Binarization methods include fixed threshold method, maximum entropy threshold segmentation method, and OTSU segmentation [37] method. In these methods, the fixed threshold method needs to manually select the threshold. The maximum entropy threshold segmentation will retain many pixels with low defect probability, which has poor robustness. This paper chooses OTSU method to segment saliency image. OTSU method is based on the idea of clustering. Firstly, the image is divided into panorama and background based on the gray value. Since variance is a measure of the uniformity of gray distribution, the greater the variance between the background and the foreground, the greater the difference between the two parts of the image. When part of the foreground is wrongly divided into background or part of the background is wrongly divided into foreground, the difference between the two parts will become smaller. Therefore, the segmentation that maximizes the variance between classes means the minimum probability of misclassification. OTSU method can highlight the regions with high defect probability in saliency map and suppress the regions with low defect probability to form binary image. The image after using OTSU method is defect segmentation image. Saliency map without defect image will be forced to set to 0, and the gray value of defect segmentation image will also be set to 0.

### 3 Experiment

#### 3.1 Image acquisition system and dataset construction

In this study, the following injector valve machine vision system is used to collect images and detect defects. This system includes Image Acquisition System, Image Analysis System, Mechanical and Electrical Control System. The Mechanical and Electrical Control System uses a Siemens S7-1500 PLC and two S7-200 PLC, as well as a number of motor control injector valves feeding and sorting. The Image Analysis System uses industrial computer with GPU for defect detection, and uses TCP/IP communication protocol to interact with PLC. This system can not only collect and analyze the surface defect image, but also detect whether the outside diameter reach standard, but these are not the focus of this paper. The image acquisition arrangement scheme is shown on the right side of Fig. 5, and the camera, lens, parallel light source and spherical light source are respectively from left to right. The parallel light source irradiates the middle part of the cylindrical valve and the spherical light source irradiates both sides of the cylindrical valve. In order to make it easy for readers to understand this system, the injector valve machine vision system is shown in Fig. 5.

There are few datasets of cylindrical metal workpiece captured by industrial cameras. This paper built a dataset of surface defects of injector valve, which is used to train and evaluate the proposed model. The dataset was collected by Basler-Aca2440-20gm camera, and the camera exposure rate was set to 2300. The size of the original image collected by the camera is $2448 \times 2048$. Since the injector valve image only accounts for a part of the image, we segmented the original image to get $768 \times 384$ valve regions, the segmentation method was obtained by using fixed threshold binarization method. Each valve region was cut into 8 images to get $192 \times 192$ valve slices as the image in the dataset. Defects often account for only a small part of the image, so the image segmentation into 8 regions is conducive to classification and generating accurate saliency map. The operation of acquiring the dataset is shown in Fig. 6. There are 6747 images in the dataset. The training dataset and testing dataset were allocated according to the 5-fold cross-validation method. Among them, pixel-level labels are included in the testing data of dirt and scratch to evaluate the segmentation accuracy. Table 2 shows the dataset distribution of training data and testing data.

#### 3.2 Evaluation metrics

We use Precision (P), Recall (R) and F-measure to evaluate classification performance. Moreover, pixel accuracy (PA) and Mean Intersection over Union (MIoU) are used to evaluate the performance of segmentation. The formula of the above evaluation indicators is as follows

$$Precision = \frac{TP}{TP + FP}$$  \hspace{1cm} (15)$$

$$Recall = \frac{TP}{TP + FN}$$  \hspace{1cm} (16)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$  \hspace{1cm} (17)$$

**Fig. 5** Injector valve machine vision system, including Mechanical and Electrical System, Image Acquisition System, Image Analysis System. The right side of figure is the hardware of acquisition system.

Pixel accuracy = \[
\frac{\sum_{i=0}^{k} P_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} P_{ij}}
\] (18)

\[MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\sum_{j=0}^{k} P_{ii}}{P_{ij} + \sum_{j=0}^{k} P_{ji} - P_{ii}}\] (19)

where TP is the number of defect images which are predicted to be correct by the model; in contrast, FP is the number of defect images which are predicted to be false. FN is the number of the actual non-defect which is mistakenly classified as defect. \(\sum_{i=0}^{k} P_{ii}\) represents the number of pixels belonging to class and predicted to be class \(i\) as well. \(\sum_{j=0}^{k} P_{ij}\) represents the number of pixels that belong to class but are predicted to be class \(j\).

Since the lack of statistical variability on the training and testing dataset. To increase the credibility of the accuracy.

**Fig. 6** Dataset building method: each image collected by camera image acquisition is divided into 8 images. These 8 images are input into the dataset.
of classification and segmentation, this paper used K-fold cross-validation to traverse all the samples to verify that the defect detection result of IRA-CNN has higher reliability. This paper used 5-fold cross-validation. The experimental steps of 5-fold cross-validation are as follows: Divide the dataset into 5 sub-samples and keep a single sub-sample as the testing dataset as well as the other 4 samples as the training dataset. The experiment was conducted for 5 times, and the average value was taken as the final test result.

### 3.3 Implementation details

The experimental platform is an industrial computer equipped with a NVIDIA GeForce RTX2080 (8G) graphics board. The CPU of the industrial computer is Intel i7-9800. The training and inference are completed on the industrial computer. The model framework is based on the Pytorch. During the experiment, we used cross entropy loss function. We use Adam optimizer to optimize the learning rate. According to [40], the initial learning rate ($\lambda$) is 0.001, the first estimated exponential decay rate ($\beta_1$) is 0.9, and the second estimated exponential decay rate ($\beta_2$) is 0.999 and the epoch of training is 100. Batch size is set to 32. In the initial state, this paper assumes that GAP and GMP have the same impact on the classification effect of the IAM. Therefore, $a$ is initially set to 0, so the weight of $F_{gap}$ and $F_{gmp}$ are initialized to 0.5. In Section 3.4.5, we compare different reduction rate. According to the experimental results, the $k$ of IAM uniformly set to 8. In the spatial attention submodule, according to [33], the kernel size of convolution layer is $7 \times 7$. In the following ablation experiments, it is shown that increase IAM can significantly improve the classification accuracy.

### 3.4 Evaluation

This subsection shows experimental results. The evaluation consists of four parts. The first part evaluates classification of injector valve defects by IRA-CNN. The second part is to evaluate the proposed segmentation framework. The third part is overall evaluation. The last part is time-efficiency evaluation.

#### 3.4.1 Classification evaluation

For evaluating classification results, since there is no article on the surface defect detection of injector valve before and the injector valve is made of stainless steel. We compare our method with metal and steel defect classification method. The artificial features of traditional metal defect classification are as follows: (1) GLCM: this method is a classical method to extract texture features. According to [41], the angle parameter is set to 0, 45, 90, and 135 and the distance parameter is set to 1, 2, 4, 8, and 16. (2) MLBP: MLBP have the advantages of rotation invariance and gray invariance. Statistical histogram of LBP feature spectrum is used as feature vector for classification. (3) HOG: the feature is formed by calculating and counting the histogram of the gradient direction of the local region of the image. HOG parameters are obtained by experimental tuning. Among them, block is set to $8 \times 8$, cell is set to $4 \times 4$, and stride is set to 4. SVM and MLR are selected in classifier, among which SVM selects linear kernel function. In addition, we compared the deep learning classification methods, such as VGG16 [38] and ResNet50. We also compared with other weakly supervised defect detection frameworks such as Decaf [39] and RWSDLDC [27], in which RWSDLDC also has attention module. We also changed IAM to CBAM for training, and added it to the result comparison. As shown in Fig. 7, the classification performance of IRA-CNN is higher than all other models in the testing dataset, reaching 97.8%. GLCM is often used to extract texture feature information; thus, it is not effective in the area where the texture feature is not obvious. the performance of classification using GLCM is the weakest. SVM classifier has good classification effect in training dataset, but its generalization ability is poor. Compared with the feature extraction method based on manual design, the deep learning method has higher classification accuracy. Compared with RWSDLDC which also has attention module, RWSDLDC only uses spatial attention module in the last feature layer, and its network depth is shallow, so its classification performance is not as good as IRA-CNN. IRA-CNN has the best classification performance for EC defects, followed by dirt and scratch, because dirt and scratch account for a small area of the image and have irregular shape. The specific data are shown in Table 3. Table 3 shows that in the 5-fold cross-validation method, the accuracy rate of each round of IRA-CNN is more than 97%, and the precision and recall of EC are greater than other defects. Precision-Recall curve of each class is drawn in Fig. 8. Figure 8 includes the PR curves of each round.

| Dataset | Defect type | Good | Total number |
|---------|-------------|------|--------------|
| (192×192) | Dirt | Scratch | EC | images | of images |
| Training | 1304 | 1464 | 1010 | 1620 | 5398 |
| Testing | 326 | 366 | 252 | 405 | 1349 |
| Total | 1630 | 1830 | 1262 | 2025 | 6747 |
| Annotation | √ | √ | × | × |
Fig. 7 Classification results of ten methods on the dataset of surface defects of injector valve.

The results showed that the AP of each class were close to 1.00. The results show that IRA-CNN has impressive classification performance.

In addition, Fig. 9 shows the confusion matrix. Similarly, the confusion matrix also contains the results of each round of 5-fold cross-validation. The confusion matrix can be used to get the TP, FP and TN of each kind of defects, so as to calculate the precision and recall. It can be observed from Fig. 10 that some Dirt defects are easily confused with scratches. These pictures are labeled as 1, 2, 3 and 4. In Fig. 10, it can be observed that the pictures labeled as 1 and 2 are actually dirt type, but they are similar to scratch type in shape. The pictures labeled 3 and 4 are of scratch type but similar in shape to the dirty type. In contrast, EC has obvious features. EC is hard to confuse with other defect types. Therefore, we conclude that labeled sample images with distinct features are crucial to achieving high classification accuracy.

### 3.4.2 Segmentation framework evaluation

According to a large number of experimental observations, EC will cover the whole injector valve, so it does not need to be segmented. Therefore, only dirt and scratch defects are selected for segmentation in this experiment. The pixel-level annotation of dirt and scratch was labeled with LabelMe. We compared our segmentation module with Ren’s and Chen’s. Chen’s SA-CAM is also a weakly supervised segmentation method, Ren’s method is similar to Chen’s, so we choose

|   | Precision | Recall | F-measure | Accuracy |
|---|-----------|--------|-----------|----------|
| 1 | 0.975     | 0.972  | 0.974     | 0.976    |
| 2 | 0.969     | 0.963  | 0.967     | 0.973    |
| 3 | 0.973     | 0.982  | 0.977     | 0.983    |
| 4 | 0.973     | 0.979  | 0.976     | 0.979    |
| 5 | 0.982     | 0.979  | 0.98      | 0.981    |

Table 3 The classification performance of IRA-CNN is evaluated by 5-fold cross-validation method.
Fig. 8  PR curve of IRA-CNN, including the results of 5-fold cross-validation method

Fig. 9  The confusion matrix of IRA-CNN, including the results of 5-fold cross-validation method

Fig. 10  Dirt surfaces labeled 1 and 2 have similar characteristics to scratch surface. Scratch surfaces labeled 3 and 4 have similar shape to dirt surface
these two methods as benchmark methods and compare it with CAM and Grad-CAM in the ablation experiment. As shown in Fig. 11 and Table 4, the area segmented by Decaf + MLR usually has a high recall rate due to its large coverage area, but the IOU value is low. Compared with SA-CAM, the precision of our segmentation framework is improved about 6.3%. This is because the segmentation method based on CAM is only sensitive to the large defect area, while our proposed framework focuses on all defect areas in the image, so it covers smaller area. Compared with the two benchmark methods, the PA and IOU of the dirt defect increased by more than 6%. The PA and IOU of scratch defect increased by more than 1.6%.

3.4.3 Overall evaluation

From above classification and segmentation evaluation of our proposed method. We can conclude that IRA-CNN is superior to other classification methods in injector valve classification, and performs well in weakly supervised defect segmentation. The reasons are as follows, IAM is integrated in every layer of IRA-CNN. This module makes

![Fig. 11 Comparison of the proposed method with Decaf + MLR and SA-CAM, the last row shows pixel-level annotation mask](image)
Table 4 The following table is the segmentation result evaluation. Including Decaf + MLR and SA-CAM and proposed method

| Method          | PA  | IOU  | Precision | Recall | F-score |
|-----------------|-----|------|-----------|--------|---------|
| Decaf + MLR     |     |      |           |        |         |
| Dirt            | 51.02 | 35.07 | 54.66     | 43.88  | 0.487   |
| Scratch         | 55.23 | 46.01 | 61.13     | 70.53  | 0.672   |
| SA-CAM          |     |      |           |        |         |
| Dirt            | 64.02 | 55.49 | 60.02     | 39.97  | 0.480   |
| Scratch         | 81.32 | 71.59 | 71.41     | 68.32  | 0.698   |
| IRA-CNN + CAM   |     |      |           |        |         |
| Dirt            | 47.41 | 39.94 | 52.95     | 42.62  | 0.478   |
| Scratch         | 63.23 | 51.21 | 65.01     | 56.79  | 0.606   |
| IRA-CNN + Grad-CAM | 63.51 | 59.04 | 65.78     | 46.31  | 0.537   |
| Scratch         | 83.25 | 71.42 | 75.69     | 70.84  | 0.717   |
| IRA-CNN + Grad-CAM++ | 70.13 | 62.94 | 68.05     | 58.77  | 0.631   |
| Scratch         | 85.47 | 73.22 | 76.17     | 76.88  | 0.765   |

IRA-CNN pay more attention to the defect area, so it improves the classification performance. IAM helps Grad-CAM++ to produce more accurate saliency map, and Grad-CAM++ makes all defect areas more prominent, which is conducive to improving the segmentation accuracy.

3.4.4 Time-efficiency evaluation

The results of FPS are calculated by testing dataset. Since the parameter of the IRA-CNN is more than Chen’s RWSLDC, the inference time is relatively slow, but compared with SVM and other machine learning classifiers, the inference speed of this method is faster. The time for each method of training and testing is given in Table 6.

3.4.5 Reduction rate of IAM

The reduction ratio $k$ introduced in Section 2.2.1 is a hyper-parameter which allows us to vary the capacity and computational cost of the IAM in the network. We set Reduction rate $K$ in IAM with different values, and the results are shown in Table 5. The results show that the parameter decreases with the increase of $K$. The accuracy is the highest when $k = 8$. Therefore, we set $K$ to 8.

3.5 Ablation studies

3.5.1 Effect of IAM

To prove the effectiveness of IAM module, the attention module is added or replaced in the ablation studies. As shown in Table 7, The effectiveness and superiority of IAM can be proved by comparing CBAM with reducing IAM. Figure 12 shows the saliency map generated by Grad-CAM++. The effect of saliency map with attention module is better than that without attention module and the saliency map generated by IAM is more accurate, the defect area is more prominent and the background area is suppressed more obviously.

3.5.2 Effect of grad-CAM++

Table 4 shows the advantages of Grad-CAM++ over Grad-CAM and CAM. In the training of CAM, we add a GAP layer to the last feature layer, and all the previous layers are frozen for retraining. The effect of CAM is the worst, because adding a GAP layer to the last layer may lose some semantic information and the classification accuracy. This can be shown in Table 4, and the Precision of CAM is greatly reduced. The effect of Grad-CAM is better than CAM, but as shown in Fig. 12, there are many dirt defects in the picture, and Grad-CAM does not cover all of them. Therefore, Grad-CAM++ has better robustness and adaptability in the task of segmentation of injector valve, especially in the task of dirty defect segmentation.

Table 5 Training time and FPS of IRA-CNN and other methods

| Reduction rate | Accuracy | Parameter |
|---------------|----------|-----------|
| 2             | 96.57%   | 17.77M    |
| 4             | 97.52%   | 17.73M    |
| 8             | 97.84%   | 17.71M    |
| 16            | 97.33%   | 17.70M    |

Table 7 Classification accuracy of different attention modules in IRA-CNN

| Method                                      | Attention                                  | Accuracy |
|---------------------------------------------|--------------------------------------------|----------|
| IRA-CNN                                     | Without attention                          | 94.53    |
| IRA-CNN                                     | IAM (without spatial attention)            | 96.25    |
| IRA-CNN                                     | CBAM                                       | 96.71    |
| IRA-CNN                                     | IAM                                        | 97.84    |
4 Conclusion

In this paper, injector valve was taken as the representative of cylindrical metal workpieces. Since the variety and complexity of cylindrical metal workpieces defects which cause pixel-level annotation requires expensive manual work. Therefore, an end-to-end weakly supervised learning framework was proposed to classify and segment surface defects. This framework only uses image tag annotation for training and performs defect classification and defect segmentation simultaneously. IAM was proposed to suppress the interference of useless background area and highlight the defect area. In addition, IRA-CNN was designed and integrated with IAM to classify the defects. The accuracy of IRA-CNN is more than 1.9% compared with other methods. The experiment is based on the 5-fold cross-validation method to verify that the accuracy can reach 97.84%. Weakly supervised learning is achieved by Grad-CAM++. In the task of segmentation, Using Grad-CAM++ to generate saliency map. The defect pixel-level segmentation based on saliency map greatly simplifies the pixel-level segmentation task of image segmentation. The segmentation precision is at least 6.3% higher than other models. The segmentation precision is improved and the labeling time is saved. This weakly supervised learning framework only needs image-level annotation. The cost of manual annotation is reduced.

There are some limits in our framework. For example, this framework cannot classify and segment unknown defects. This is a common problem in industrial environment. In the future, we will study how to integrate unclassified defects into classification networks and realize segmentation and continue to optimize our IRA-CNN, and try to use different convolutional kernel forms to improve classification accuracy.

Acknowledgements The authors would like to express their appreciation to the developers of the Pytorch and the developers of Grad-CAM and Grad-CAM++.

Author contribution M. Ye designed the methodology and conducted the experiment under the guidance of W.W. Zhang, X.L. Wang and G.H. Cui undertook data check and investigation and acquired resources.

Funding This work was supported in part by the National Natural Science Foundation of China (No. 51805312) and in part by Shanghai Sailing Program (No. 18YF1409400).

Declarations

Ethics approval The authors declare that this manuscript was not submitted to more than one journal for simultaneous consideration. Also, the submitted work was original and has not been published elsewhere in any form or language.

Consent to participate Not applicable
References

1. Malamas EN, Petrakis E, Zervakis M, Petit L, Legat JD (2003) A survey on industrial vision systems, applications and tools. Image Vis Comput 21:171–188
2. Hanzaei SH, Afshar A, Barazande F (2017) Automatic detection and classification of the ceramic tiles’ surface defects. Pattern Recogn 66:173–189
3. Park Y, Kweon IS (2016) Ambiguous surface defect image classification of amoled displays in smartphones. IEEE Trans Indust Inform 12:597–607
4. Song K, Yan Y (2013) A noise robust method based on completed local binary patterns for hot-rolled steel strip surface. Appl Surf Sci 285:858–864
5. Apostolos C, Ivan P, Ivan J (2016) Feature selection for surface defect classification of extruded aluminum profiles. Int J Adv Manuf Technol 83:33–41
6. Shumin D, Zhoufeng L, Chunlei L (2011) AdaBoost learning for fabric defect detection based on HOG and SVM. In: 2011 IEEE International conference on multimedia technology, pp 2903–2906
7. Kwon BK, Won JS, Kang DJ (2015) Fast defect detection for various types of surfaces using random forest with VOV features. Int J Precis Eng Manuf 16:965–970
8. Han H, Gao C, Zhao Y, Liao S, Tang L, Li X (2020) Polycrystalline silicon wafer wafer defect segmentation based on deep convolutional neural networks. Pattern Recognit Lett 130:234–241
9. Augustauskas R, Lipnickas A (2020) Improved Pixel-Level Pavement-Defect segmentation using a deep autoencoder. Sensors 20(9):2557
10. Baumgartl H, Tomas J, Buettner R, Merkel M (2020) A deep learning-based model for defect detection in laser-powder bed fusion using in-situ thermographic monitoring. Progress Addit Manuf 1–9
11. Xu X, Zheng H, Guo Z, Wu X, Zheng Z (2019) SDD-CNN: Small Data-driven convolutional neural networks for subtle roller defect inspection. Appl Sci 9(7):1364
12. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
13. Chen H, Pang Y, Hu Q, Liu K (2020) Solar cell surface defect inspection based on multispectral convolutional neural network. J Intell Manuf 31(2):453–468
14. Cheon S, Lee H, Kim CO, Lee SH (2019) Convolutional neural network for wafer surface defect classification and the detection of unknown defect class. IEEE Trans Semiconductor Manuf 32(2):163–170
15. He Y, Song K, Dong H, Yan Y (2019) Semi-supervised defect classification of steel surface based on multi-training and generative adversarial network. Opt Lasers Eng 122:294–302
16. Li J, Su Z, Geng J, Yin Y (2018) Real-time detection of steel strip surface defects based on improved yolo detection network. IFAC-PapersOnLine 51(21):76–81
17. Redmon J, Farhadi A (2018) YoloV3: An incremental improvement. arXiv:1804.02767
18. Su B, Chen H, Zhou Z (2020) BAF-Detector: An Efficient CNN-Based Detector for Photovoltaic Solar Cell Defect Detection. arXiv:2012.10631
19. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv:1506.01497
20. Domen T, Šela S, Skvarč J, Skočaj D (2020) Segmentation-based deep-learning approach for surface-defect detection. J Intell Manuf 31(3):759–776
21. Chen L, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 801–818
22. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp 234–241
23. Tao X, Zhang D, Ma W, Liu X, Xu D (2018) Automatic metallic surface defect detection and recognition with convolutional neural networks. Appl Sci 8(9):1575
24. Wang M, Cheng JCP (2020) A unified convolutional neural network integrated with conditional random field for pipe defect segmentation. Comput-Aided Civ Infrastruct Eng 35(2):162–177
25. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929
26. Lin H, Li B, Wang X, Shu Y, Niu S (2019) Automated defect inspection of LED chip using deep convolutional neural network. J Intell Manuf 30(6):2525–2534
27. Chen H, Hu Q, Zhai B, Chen H, Liu K (2020) A robust weakly supervised learning of deep Conv-Nets for surface defect inspection. Neural Comput Appl 1–16
28. Xu L, Lv S, Deng Y, Li X (2020) A weakly supervised surface defect detection based on convolutional neural network. IEEE Access 8:42285–42296
29. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
30. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp 839–847
31. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
32. Ji Y, Zhang H, Wu QMJ (2018) Salient object detection via gradient-based localization. In: Proceedings of the IEEE conference on computer vision and pattern recognitionm, pp 4510–4520
33. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
34. Zhao J, Shao J, Shao H, Wang Y, Liu D (2019) Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp 839–847
35. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
36. Ji Y, Zhang H, Wu QMJ (2018) Salient object detection via multi-scale attention CNN. Neurocomputing 322:130–140
37. Woo S, Park J, Lee JY, Kweon IS (2018) CBAM: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
38. Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning, pp 6105–6114
39. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
40. Lin M, Chen Q, Yan S (2013) Network in network. arXiv:1312.4004
41. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9(1):62–66
38. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
39. Ren R, Hung T, Tan KC (2017) A generic deep-learning-based approach for automated surface inspection. IEEE Trans Cybern 48(3):929–940
40. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
41. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution grayscale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.