[1,*]Dayong Yu
[2]Shuhui Yang

Defect Detection of Cell Phone Screen Using a Faster Regional Convolutional Neural Network with Multi-head Attention Mechanism

Journal of Electrical Systems

Abstract: - Since the glass outer screen of a cell phone is the main sensory part of the human eye when using a cell phone, the advantages and disadvantages of the cell phone screen directly affect people's sense of use. Therefore, the defect detection requirements for cell phone screens are high and need to meet the needs of high-volume factory inspection. Most of the traditional defect detection methods use visual methods, the detection results are overly dependent on the subjectivity and experience of workers, the efficiency of this method is low, and the accuracy is poor. Currently, machine learning-based detection methods are applied in numerous industries. In this paper, a faster Regional Convolutional Neural Network (R-CNN) with multi-head attention mechanism for defect detection of cell phone screen is proposed. To enhance the network's capability in extracting feature information, a four-head attention mechanism is added to the last convolutional layer of the ResNet50 network. An improved Region of Interest (ROI) Align is proposed to replace the original ROI Pooling to reduce the localization error of cell phone screen defects. Replace the original Rectified Linear Unit (ReLU) activation function with the Copy Exponential Linear Unit (CELU) activation function to expedite the convergence capability of the network. Finally, by comparing with other classical model training, the evaluation results indicate that the proposed method achieved an average accuracy of 95.71%, which is a 5.34% improvement compared to the original faster R-CNN network.

Keywords: Defect Detection (DD), Convolutional Neural Network (CNN), Image Recognition (IR), Multi-Head Attention Mechanism (MHAM), Cell Phone Screen (CPS).

I. INTRODUCTION

In recent years, deep learning-based object detection algorithms have been widely applied in various fields such as facial recognition [1], aerospace, medical diagnosis, and intelligent surveillance. Many scientists work on machine learning-based defect detection methods [2]. Weimer [3] et al proposed a deep convolutional neural network (DCNN) technique for defect detection in industry, which automatically generates robust features from amounts of training data with minimal expert knowledge through a hierarchical learning strategy. The results show that the method of deep convolutional neural networks outperforms the currently available techniques in terms of overall detection accuracy. Li [4] and his colleagues proposed a defect detection technique that combines multilayer perceptron with deep learning. Additionally, it utilizes a coarse-precision strategy to improve the efficiency of the model. The results show that this algorithm performs well in detecting defects such as scratches, foreign objects, and light stains. Kuchipudi [5] et al. proposed a region-based CNN for automatically detecting, localizing, and segmenting defects in noisy ultrasound images corresponding to multiple features. Moreover, this network utilizes ROI Align instead of traditional ROI Pooling. Compared with several state-of-the-art defect detection networks, the results show that this network achieved the best mean average precision (mAP) of 0.98 on the test set. Fu [6] et al. used faster R-CNN to train, validate, and realize the automatic identification of pavement cracks and compared it with the automatic identification method of U-Net segmentation of pavement cracks. The result shows that the recall and accuracy are greatly improved by the proposed detection method with deep learning, and both achieve more than 85% results. The results of the faster R-CNN method are closer to the real condition of pavement cracks than the U-Net segmentation automatic identification method. Mansour [7] et al. combined Faster R-CNN with deep reinforcement learning models for video anomaly detection and classification. The paper uses a deep Q learning based reinforcement learning model to classify the detected anomalies. The final accuracy on datasets Test004 and Test007 reaches 98.5% and 94.8%, respectively. Chen [8] et al. utilized the MobileNet network in conjunction with the SSD model for defect detection in curved parts, achieving over 90% recognition accuracy for two defect categories in the dataset. Additionally, they achieved a detection speed of 0.78 seconds per image, making a significant contribution to defect detection in the ceramic industry as a whole. In summary, the detection method using deep learning has the advantages of higher accuracy and higher efficiency. However, compared with

[1] School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai, China
[2] School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai, China
*Corresponding author: Dayong Yu

road cracks and other surface defects, the defects of cell phone screens are relatively small. Therefore, the detection of cell phone defects has relatively high requirements on the network for feature extraction.

Therefore, to address these issues, this paper added a multi-head attention mechanism in Faster R-CNN to represent the problem of incomplete feature extraction during the feature extraction process. The network replaces ROI Pooling with ROI Align on the original network structure, thus reducing the localization error. The research results display that the average accuracy of this method can be as high as 95.71%, which is improved by 5.34% on the original network.

II. RELATED WORK

A. Data Sets and Image Preprocessing

1) Data set acquisition

In this paper, 400 sheets of each of the three types of defects, stains, scratches, and dirty were collected using an industrial camera. Smudges are pits with small areas and a certain depth on the surface of the glass screen due to improper operation of the engraving head, grinding head, etc. during the opening, slotting, chamfering, and flat grinding processes. The image of the defect is shown in Figure 1.

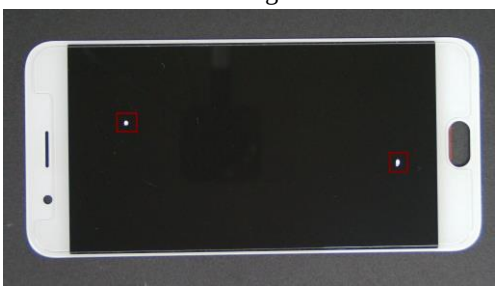

Figure 1: Stains Defect

Scratches are mainly caused by a narrow indentation on the surface of the glass screen due to improper operation of the grinding head or flat grinding equipment during the grooving, chamfering, and flat grinding process. The length is about 15mm, and the scratch defect is shown in Figure 2.

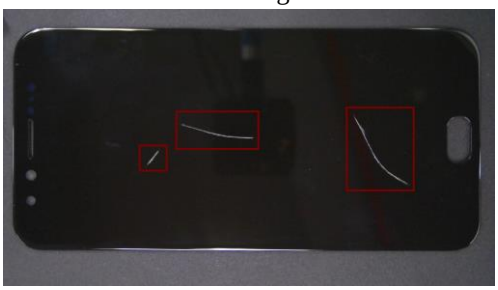

Figure 2: Scratches Defect

Dirt is mainly in the printing, it is possible that the screen leakage of ink improperly, resulting in a certain area of ink in the glass sheet blank without the need for inking. The defects are shown in Figure 3, with a length of about 20mm and a width of about 6mm.

Figure 3: Dirty Defect

Due to the training in deep learning, the larger amount of data, the higher accuracy of network learning. Therefore, in this paper, 4800 defect images are finally obtained utilizing data augmentation [9], including up-down and left-right flipping and clockwise rotation of ninety degrees. Flipping usually refers to flipping in the horizontal or vertical direction, in this paper, the horizontal flipping method is used, and the defective image of the cell phone screen takes the horizontal axis of symmetry as the axis, and the double defective samples of the cell phone screen are generated by the flipping operation. The rotation operation is similar to the flip operation, the

rotation center is usually the geometric center of the image center, and the image is randomly rotated within a certain rotation angle. In this paper, the rotation angle is set to 30° clockwise, and the data labels are preserved. Figure 4 shows the effect of using vertical and horizontal flipping as well as clockwise rotation in this paper.

Finally, the defect types and bounding box locations in the cell phone cover glass screen images are labeled according to the VOC2007 dataset format. Divide all defect images into training sets, validation sets, and test sets in a ratio of 7:1.5:1.5. The following Table 1 shows the data distribution of each kind of defect on the dataset.

Table 1: Distribution of Data Sets

	Train	Validation	Test	Total
Dot	1120	240	240	1600
Line	1120	240	240	1600
Block	1120	240	240	1600
Total	3360	720	720	4800

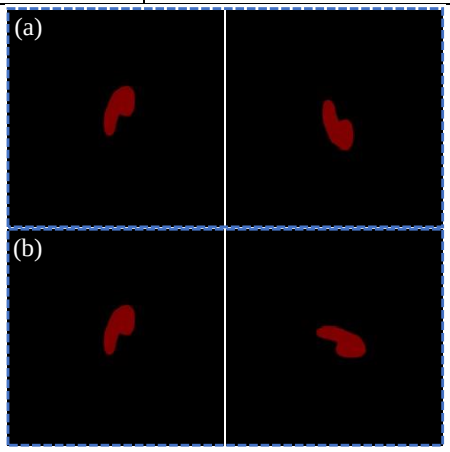

Figure 4: Data Enhancement (a) Turn upside-down; (b) Turn clockwise

2) Preprocessing algorithms

In industrial production, most of the cell phone screen images we obtain are with redundant backgrounds and great noise, because of the great influence on the subsequent experimental analysis and calculation. Therefore, preprocessing of images is crucial, as it aims to eliminate irrelevant features in the images, retain useful information, and enhance detectability. The preprocessing steps for the research content of this paper are as follows: grayscaling, filtering and denoising, image segmentation, and geometric transformation.

The operation of converting a color image into a grayscale image is called image grayscaling. For RGB images, which are gray-scale images at that time, grayscale images occupy less memory and faster computing speed. Additionally, converting to grayscale images can visually enhance contrast and highlight target areas. This paper uses a weighted averaging method for grayscale conversion, the calculation principle is shown in the following formula:

$$I(x,y) = 0.299 * I_R(x,y) + 0.578 * I_G(x,y) + 0.114 * I_B(x,y) \quad (1)$$

The purpose of filtering and denoising is mainly to remove the interference of noise to avoid interference in the subsequent image analysis, the most commonly used in image processing is the median filtering method, but due to the use of median filtering on the edges of the cell phone screen denoising effect is not particularly ideal, such as Figure 5(a), the edge of the phone burr is more serious, the edge information is not ideal, so we use the adaptive median filtering, such as Figure 5(b), and find that it can better remove noise and preserve edge information. It is found that the noise can be removed better and the edge information is preserved. Adaptive filtering is currently one of the best noise reduction methods, as it offers superior adaptability and filtering capabilities. The principle of adaptive median filtering is to dynamically adjust the window size of the median filter based on preset conditions, balancing the effects of noise reduction and edge detail preservation.

In image processing, it is often necessary to extract the Region of Interest (ROI) to simplify the workflow of this paper. Simply put, the region of interest is selected from an image, and the cell phone screen region is the main focus of this paper. In this paper, this region is selected for further processing. The region of interest can not only reduce the computational load of the network but also decrease processing time, thus enhancing work efficiency. There are many types of common edge detection operators, but the Canny operator has the best noise suppression effect and can detect the edge information well.

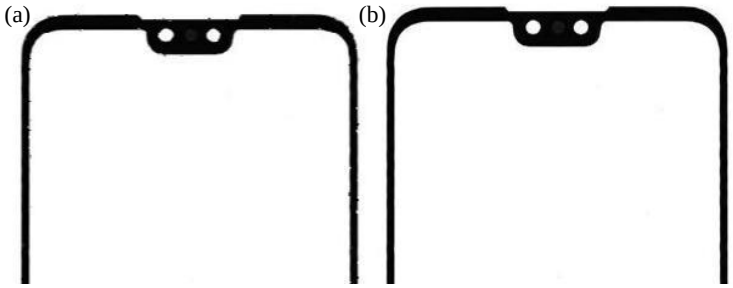

Figure 5: Filtering and Denoising Methods. (a) Median filter; (b) Adaptive median filter

Therefore, in this paper, the Canny operator is chosen for edge detection, and the edge details are optimized by combining the morphological closure operation, and the final edge image obtained is shown in Figure 6.

Figure 6: Edge Detection Results

B. Original Neural Network Structure

The Faster R-CNN algorithm is an extension of R-CNN [10] and Fast R-CNN [11], as shown in Figure 7. As the name suggests, the function of a feature extraction network is to extract the feature information from input images. These features are then input to the Region Proposal Network (RPN), which uses a sliding window combined with different scales and ratios to generate proposed target regions [12]. Subsequently, the outputs of both networks are simultaneously input into ROI Pooling to generate feature maps of a fixed size. Finally, pass the generated feature maps to a fully connected layer for classification and regression tasks.

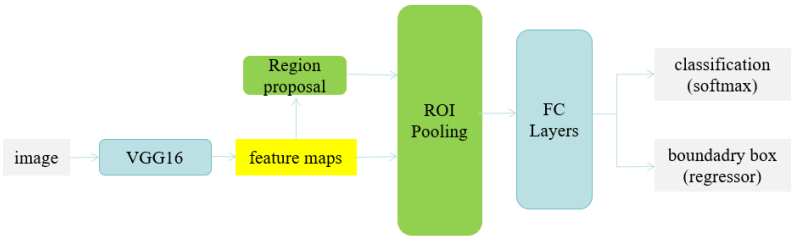

Figure 7: Faster-RCNN Network Structure

III. IMPROVING THE ALGORITHM

A. Overall Framework

The feature extraction network used in the original Faster R-CNN is VGG16, but VGG16 is difficult to ensure that the defective features on the cover glass surface are adequately extracted due to its limited number of convolutional layers (13 layers), and poor detection of small targets. However, using deeper networks to enhance feature extraction capabilities, problems such as gradient vanishing and overfitting may occur [13]. In contrast, the residual network ResNet can effectively avoid the above problems and extract more feature information of small targets due to the inclusion of residual module structure in the stacked convolutional layers. In this paper, after testing ResNet networks with different depths, ResNet50 is chosen as the feature extraction network. ResNet50 uses jump connections, and its inputs and outputs are directly connected, as shown in Figure 8, where the two 1×1 convolutions are for dimensionality reduction and dimensionality upgrading, and the 3×3 convolution is for extracting feature information. And since ROI Pooling in the original structure introduces two quantization operations, it can lead to the region mismatch problem. Therefore, replacing ROI Pooling with ROI Align can avoid quantization errors and reduce the localization errors of defects on the glass outer screen. The improved network structure proposed is shown in Figure 9.

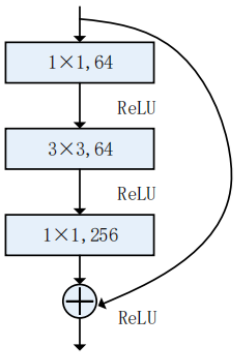

Figure 8: Residual Structure

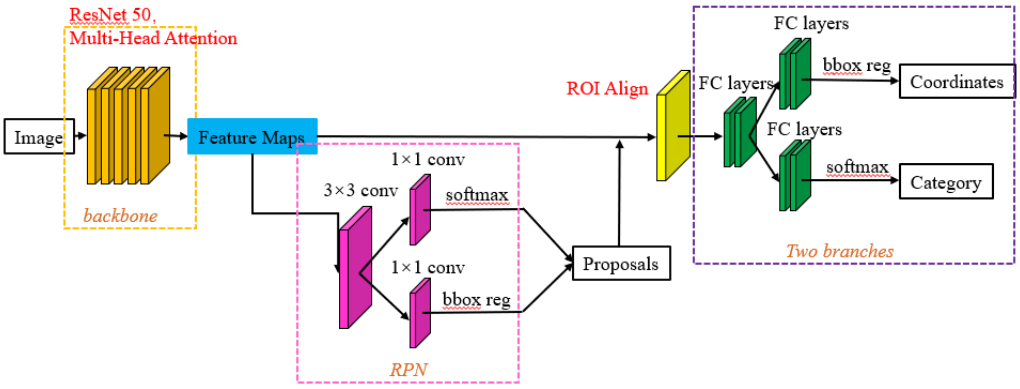

Figure 9: Improved Structure of Faster-RCNN

B. Detailed Improvements

1) Multi-attention mechanisms

In recent years, the attention mechanism has achieved remarkable advancements in image processing, natural language processing, and object detection, clearly demonstrating its ability to improve model performance. Introducing the attention mechanism into the network can enhance the feature extraction from images while increasing the weights of features related to defects [14]. Attention mechanisms apply corresponding attention weights to image features, highlighting important features of the target object and suppressing irrelevant information. Common attention mechanisms currently used include self-attention mechanism, multi-head attention mechanism [15], spatial attention mechanism [16], and channel attention mechanism [17]. The multi-head attention mechanism is a more complex version of the self-attention mechanism, representing a variant of self-attention [18], aiming to enhance the expressive and generalization capabilities of the model. Figure 10 is a diagram of a multi-head attention mechanism, where Q, K, and V are three fixed values mapped through Linear layers. These values are processed by the Scaled Dot-Product Attention scoring function and concatenated with the output of each head. Finally, they are mapped back to the output similar to that of a single attention head through Linear transformation. Each head represents a different type of attention, selecting distinct information. By utilizing several individual attention heads to compute attention weights independently and combining all results through concatenation or weighted sum, a richer representation is obtained. The multi-head attention mechanism achieves this by parallel linear transformations and attention calculations for multiple sets of queries, keys, and values, and eventually connecting them to enhance the model's focus on different features. This helps the model to capture the information in the input data in a more comprehensive way, allowing the model to achieve better results. Therefore, this paper defines the input of the multi-head attention mechanism as the final output of ResNet50's last convolutional layer, enhancing the capability of feature information extraction. The improved ResNet50 network structure is shown in Figure 11. The formula for the multiple attention mechanism is as follows:

$$Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V, i = 1,...,4 \tag{2}$$

$$head_i = Attention\left(Q_i, W_i, V_i\right), i = 1,...,4 \tag{3}$$

$$MultiHead\left(Q,W,V\right) = Contact\left(head_1,...,head_4\right)W^o \tag{4}$$

Where Q is the query matrix, K is the key matrix and V is the value matrix.

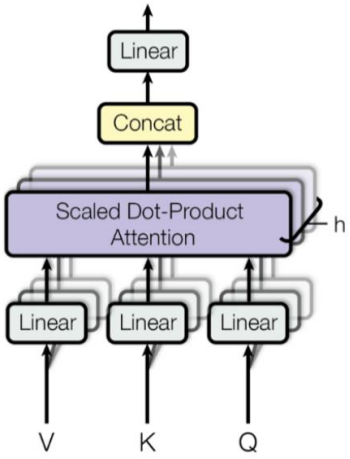

Figure 10: Structure of the Multi-attention Mechanism

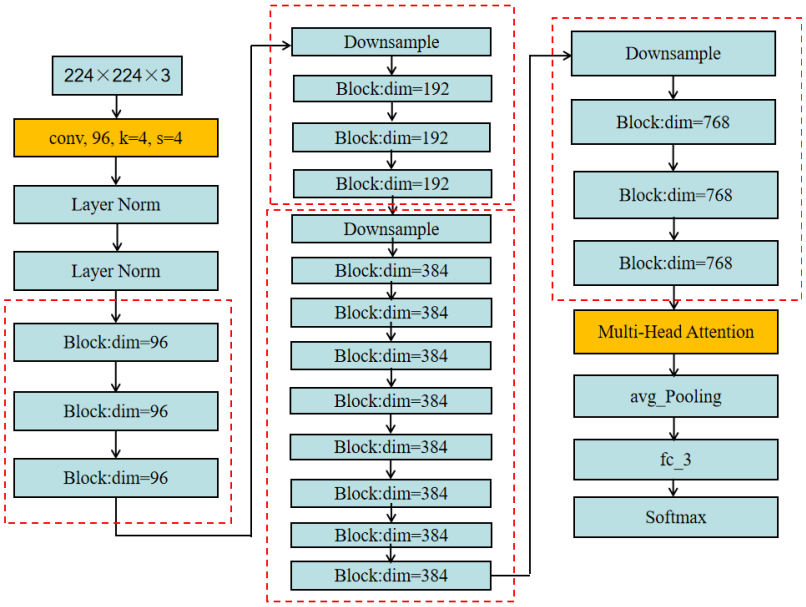

Figure 11: Improved Residual Network Structure

2) ROI align

ROI Pooling is done by dividing the region of interest from the feature map and then making it a fixed size and dimension feature map. However, the feature mapping and chunking process can lead to region mismatch problems due to the introduction of two quantization operations for rounding. In order to prevent quantization errors and enhance localization accuracy, this paper opts for ROI Align over ROI Pooling to address the problem of region mismatch. The operation flow of ROI Align is shown in Figure 12. First, each candidate region is traversed, second, the candidate regions are equally divided into m×m cells (m = 2 in the figure). After being equally divided, the vertices are unlikely to fall on real pixels. Therefore, four fixed pixels are selected for value calculation, where the value at each point is interpolated bilinearly based on the values of the nearest four actual pixels. Finally, the maximum value among them is taken as the resulting value, with an output size of m×m, where m is set to 2. The improvement of ROI Align can obtain more accurate ROI location information of glass cover defects, thus improving defect detection accuracy.

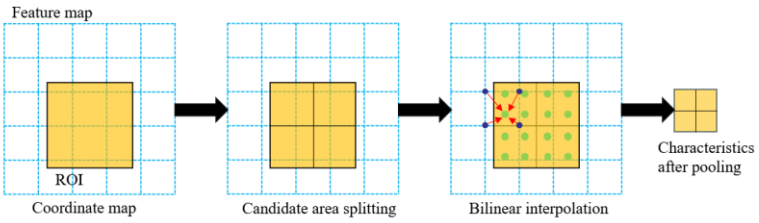

Figure 12: ROI Align

3) Activation functions

ReLU is an activation function widely used in neural networks and is known for its fast computation and powerful performance. Still, the output of the function is zero for input x <0, and the loss gradient disappears during backpropagation, which results in the parameters not being able to be updated, resulting in neuron death. To address this issue, this paper improves the ResNet50 network by selecting CELU [19] as the activation function. CELU is a nonlinear function with a kink that also possesses continuity and differentiability, which can benefit the convergence of neural networks. The CELU activation function is calculated as follows, taking α as 0.075, and the output value of the function is compared as in Figure 13.

$$CELU(x,\alpha)=\begin{cases} x & if\ x\geq 0 \\ \alpha\left(\exp\left(\frac{x}{\alpha}\right)-1\right), if\ x<0 \end{cases} \tag{5}$$

Kaiming normal distribution in the CELU activation function can accelerate the convergence of the model with better results [20], the model in this paper chooses Kaiming normal distribution as the initialization weight way.

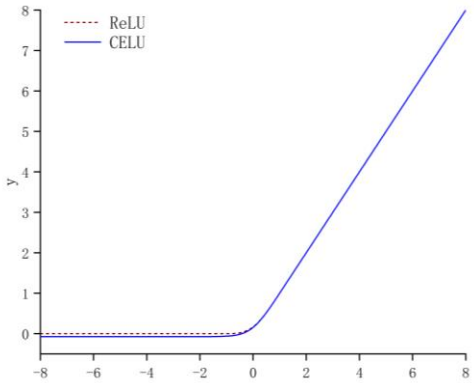

Figure 13: Comparison Curve of Activation Functions ReLU and CELU

IV. EXPERIMENTAL COMPONENT

A. Experimental Configuration and Computer Configuration

The hardware environment and main software configurations used in the experiment are shown in Table 2. During the model training process, Faster R-CNN is trained using stochastic gradient descent optimization algorithm with momentum (SGDM), and the learning rate is set to 0.0002, momentum to be 0.9, weight decay to be 0.0001, Epoch to be 50 times, Batch Size=5, and the image size to be 224×224.

Table 2: Experimental Environment and Software Configuration

Name	Parameters
CPU	Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz
GPU	Geforce RTX 4050
Operating system	Windows10
Editing software	MATLAB 2020b

B. Assessment Criteria

Typically, the false detection rate and leakage rate are common metrics used to assess the accuracy of defect detection. The false detection rate is generally measured by mean average precision (mAP); the leakage rate is generally measured using the recall rate (Recall). When the average accuracy rate rises, the false alarm rate tends to decrease; similarly, an increase in the recall rate leads to a decrease in the miss detection rate. The recall rate is calculated based on the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for the four classes, as detailed in Table 3.

Table 3: Confusion Matrix

Category		Predictive labeling		Total
		0(no defect)	1(defect)	
truthful labeling	0(no defect)	*TP*	*FN*	*P*
	1(defect)	*FP*	*TN*	*N*
	Total	*P`*	*N`*	*P+N*

Recall is a measure that evaluates the percentage of accurately predicted positive samples among all positive samples, and it is calculated as follows:

$$\text{Re}call = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{6}$$

The Average Precision (AP) is determined by computing the area under the Precision-Recall curve, while mAP denotes the average of AP across each image category [21]. mAP is calculated using the formula:

$$mAP = \frac{\sum_{i=1}^{n} \int_0^1 P(R)dR(n)}{N} \times 100\% \tag{7}$$

Where: P(R) is the Precision-Recall curve, n is the category number, and N is the number of categories.

C. Comparative Tests and Results

1) Comparison test of different feature extraction networks

To assess the effectiveness of the enhanced ResNet50 network introduced in this paper, four networks, AlexNet, MobileNetV2, ResNet18, VGG16, and ResNet50, are selected in this paper to be compared with the improved ResNet network as the feature extraction network for Faster R-CNN. The detailed experimental results are presented in Table 4.

Table 4: Experimental Results of Different Feature Extraction Networks

Model	mAP/%	Recall/%
AlexNet	77.80	75.45
MobileNetV2	83.92	81.24
ResNet18	86.73	83.51
VGG16	88.65	86.30
ResNet50	88.90	86.92
Improved ResNet50	89.34	87.67

From Tab. 4, it is evident that when using the improved residual network proposed in this paper to build the faster R-CNN, the feature extraction capability is significantly enhanced the detection effect is improved significantly, the average precision reaches 89.34% and the recall rate reaches 87.67%. Compared to other networks, the improved ReaNet50 network has the highest average precision and recall, i.e., the lowest leakage and false detection rates, demonstrating strong feature extraction capabilities. Compared to AlexNet, MobileNetV2, ResNet18, VGG16, and ResNet50 in terms of average precision, it improves 11.54%, 5.42%, 2.61%, and 0.69%, 0.44%, and recall improves 11.22%, 5.43%, 3.16%, and 1.37%, 0.75%, respectively. The improved network proposed in this paper effectively enhances the network's ability to extract feature information and significantly reduces the leakage and false detection rates due to the adoption of the multi-head attention mechanism.

2) Classical network comparison test

To further confirm the effectiveness of the enhanced network proposed in this paper, YOLOv2, SSD, Faster R-CNN, and the proposed network were chosen for comparison. The results of the experiments are displayed in Table 5.

Table 5: Comparison between the Model of This Paper and the Classical Model

Model	mAP/%	Recall/%
YOLOv2	80.12	75.61
SSD	85.98	79.86
Faster R-CNN	90.37	87.25
improved-Net	95.71	89.15

From the results, the improved network proposed has a great reduction in the leakage rate and false detection rate compared to other classical models, and the speed has been shortened. The ultimate training outcome of the proposed network model in this paper attains an average precision of 95.71% and a recall of 89.15%. Compared to YOLOv2, SSD, and the original Faster R-CNN network, the average accuracy is improved by 15.59%, 9.73%, and 5.34%, respectively. In recall, it is improved by 13.54%, 9.29%, and 1.9%, respectively. Thus, it can be shown that the improved network in this paper is effective and can fulfill the detection needs of the industry.

3) Comparison of model detection effectiveness

To continue to verify the detection effect of the improved network, three images of glass cover defects containing different types of defects are selected for testing and verification, as shown in Fig. 14. Fig. 14(a) shows the image to be detected, in which the defective target to be detected is labeled with a dashed box, and Fig. 14(b)-(d) show the detection results of the three original Faster-R-CNN, YOLOv2, and improved detection models, respectively.

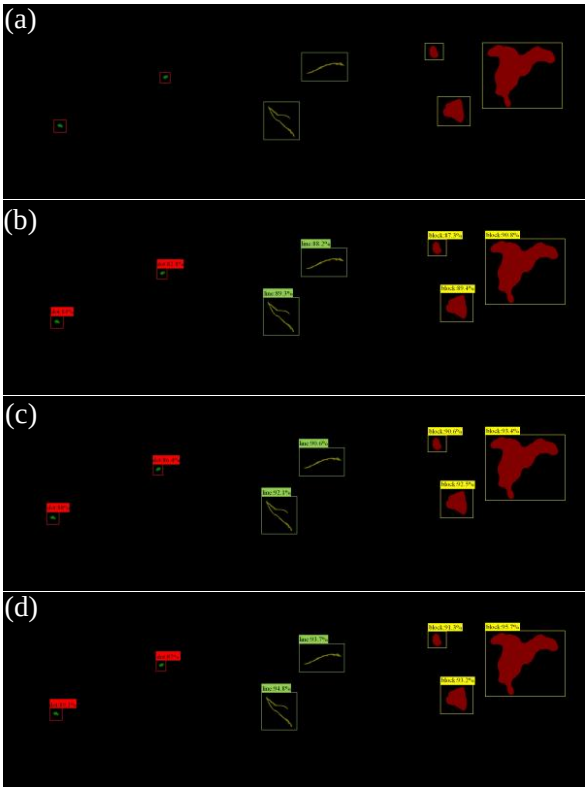

Figure 14: Test Results: (a) Image to be detected (b) YOLOv2 algorithm (c) Faster-R-CNN algorithm (d) Improved network algorithm

Observing Figure 14, it is apparent that when utilizing the enhanced network suggested in this paper to detect defects on mobile phone screens, there is a significant improvement in detecting different types of defects compared to other classical networks, and there is also an improvement in detection time. Therefore, it can be shown that the improved network algorithm in this paper can more accurately identify the defect categories and locations of cell phone glass covers, with higher average precision and recall, and better solve the problems of leakage and misdetection.

V. CONCLUSIONS

In this paper, MHFCA-Net is proposed to realize the detection of different types of defects for cell phone screens, which achieves 95.71% accuracy and 89.15% recall on a home-made dataset, which is a good detection result and can be used as an alternative to manual detection. The average accuracy of the model proposed also reaches 90.67% for the input detection of a single image. However, the average accuracy is 86.1% when using the YOLOv2 network, and the original Faster R-CNN network achieves an average detection accuracy of 89.2%. This fully demonstrates the excellence and accuracy of the improved model. In addition, the improved ResNet50 network incorporates a dual-head attention mechanism to enhance the information extraction ability of feature information. Furthermore, employing ROI Align eliminates the need for quantization in ROI Pooling, thereby decreasing localization errors for defects with extreme aspect ratios. Although the improved model in this thesis achieves higher accuracy in defect detection, it still has shortcomings. There are still small oscillations in the accuracy and loss function change curves in the late stage of model training, and the convergence speed is slow. Therefore, in the follow-up work, further parameter optimization will be carried out to achieve convergence.

In future work, the following research will be conducted: (1) extending the application of the proposed method to a broader spectrum of mobile phone defect detection categories; (2) further exploring optimization algorithms to enhance detection speed.

DISCLOSURES

The authors declare that there is no conflict of interest.

CODE, DATA, AND MATERIALS AVAILABILITY

Data are available from the authors upon request.

REFERENCES

[1] X. Sun, P. Wu, S. C. Hoi, Face detection using deep learning: an improved faster rcnn approach, Neurocomputing, vol. 299, pp. 42-50, 2018.

[2] S. B. Jha and R. F. Babiceanu, Deep CNN-based visual defect detection: Survey of current literature, Computers in Industry, vol. 148, pp. 103911, June 2023.

[3] D. Weimer, B. Scholz-Reiter and M. Shpitalni, Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection, CIRP Annals, vol. 65, pp. 417–420, 2016.

[4] C. Li, X. Zhang, Y. Huang, C. Tang and S. Fatikow, A novel algorithm for defect extraction and classification of mobile phone screen based on machine vision, Computers & Industrial Engineering, vol. 146, pp. 106530, August 2020.

[5] S. T. Kuchipudi and D. Ghosh, Automated detection and segmentation of internal defects in reinforced concrete using deep learning on ultrasonic images, Construction and Building Materials, vol. 411, pp. 134491, January 2024.

[6] Q. Fu, F. Pu, H. Ren and J. Gong, Target detection of pavement cracks based on deep learning methods, Highway, vol. 68, pp. 395–405, 2023.

[7] R. F.Mansour, J. Escorcia-Gutierrez, M. Gamarra, J. A.Villanueva and N. Leal, Intelligent video anomaly detection and classification using faster RCNN with deep reinforcement learning model, Image and Vision Computing, vol. 112, pp. 104229, August 2021.

[8] W. Chen, B. Zou, C. Huang, J. Yang, L. Li, J. Liu and X. Wang, The defect detection of 3D-printed ceramic curved surface parts with low contrast based on deep learning, Ceramics International, vol. 49, pp. 2881–2893, January 2023.

[9] Z. Shi and S. Lei, A review of image super-resolution reconstruction algorithms, Data Acquisition and Processing, vol. 35, pp. 1–20, 2020.

[10] C. Tao, C. Cao, H. Cheng, Z. Gao, X. Luo, Z. Zhang and S. Zheng, An efficient 3D detection method based on Fast Guided Anchor Stereo RCNN, Advanced Engineering Informatics, vol. 57, pp. 102069, August 2023.

[11] M.Xin, Y. Wang and X. Suo, Based on Fast-RCNN multi target detection of crop diseases and pests in natural light, Springer, Cham, vol. 81, pp. 132–139, July 2021.

[12] B. Hu and J. Wang, Detection of PCB Surface Defects With Improved Faster-RCNN and Feature Pyramid Network, IEEE Access, vol. 8, pp. 108335–108345, 2020.

[13] M. Sha, Y. Li and A. Li, Improved Faster R-CNN for multiscale aircraft target detection in remote sensing images, Journal of Remote Sensing, vol. 26, pp. 1624–1635, 2022.

[14] D. Mrowca, M. Rohrbach, J. Hoffman, R. Hu, K. Saenko and T. Darrell, Spatial semantic regularisation for large scale object detection, 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1003–1011, 2015.

[15] H. K. Shin, Y. H. Ahn, S. H. Lee, H. Y. Kim, Automatic concrete damage recognition using multi-level attention convolutional neural network, Materials, vol. 13, pp. 5549, 2020.

[16] M. Cong, C. Lu, D. Liu, Q. Xiao and R. Li, Refine-ACTDD-based method for detecting minor defects in the appearance of castings, Computer-integrated Manufacturing System, vol. 28, pp. 2815–2824, 2022.

[17] M. Jederberg, K. Simonyan, A. Zisserman and K. Kavukcuoglu, Spatial transformer networks, Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 2017–2025, December 2015.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N.Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, Computer Science, June 2017.

[19] J. T.Barron, Continuously Differentiable Exponential Linear Units, ArXiv, April 2017.

[20] T. Henmi, E. R. R. Zara, Y. Hirohashi and T. Kato, Adaptive signal variances: CNN initialization through modern architectures, 2021 IEEE International Conference on Image Processing (ICIP), pp. 374–378, September 2021.

[21] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int J Comput Vision, vol. 88, pp. 303-338, 2010.