The International Journal of Advanced Manufacturing Technology (2023) 127:5817–5829
https://doi.org/10.1007/s00170-023-11443-9

APPLICATION

Surface defect detection of smartphone glass based on deep learning

Yuechu Mao[1,2] · Julong Yuan[1,2] · Yongjian Zhu[3,4] · Yingguang Jiang[1,2]

Received: 16 November 2022 / Accepted: 15 April 2023 / Published online: 6 July 2023

Abstract

Because of the high difficulty of smartphone glass detection and the variety of defect morphologies, the detection results are easily affected by the environment, making it difficult to meet the accuracy requirements of industrial inspection. Based on the existing YOLO v5s network, this study proposes a new network Dy-YOLO v5s. In particular, an attention module is introduced into the residual structure, and the cross-scale and cross-layer connections of feature maps are added to the Neck to improve the feature extraction and information exchange capabilities of the detection network. This algorithm introduces the dynamic detection framework called dynamic head (DyHead), which improves the detection head's capacity for perception. Additionally, the redundant anchor boxes and the balance of positive and negative samples are deduplicated using the confidence propagation cluster (cp-cluster) and varifocal loss functions. The experimental results demonstrate that when the intersection over union (IOU) threshold is set to 50%, the mean average precision (mAP) of Dy-YOLO v5s, precision rate (P), and recall rate (R) reach values of 96.2%, 92.6%, and 93.1%, respectively. Compared with YOLO v5s, mAP@0.5 and mAP@0.5−0.95 increased by 4.5% and 4.6%, respectively. The approach also has significant advantages over other deep-learning algorithms in terms of overall accuracy and real-time performance. Therefore, it can fully satisfy the detection requirements of smartphone glass.

Keywords Deep learning · YOLO · Smartphone glass · Defect inspection · Defect detection · Computer vision

1 Introduction

The annual growth rate of the global smartphone industry and the advent of the 5G era have both contributed to an increase in the volume of smartphone transactions recently. Smartphones have become an indispensable necessity for people's work, life, leisure, entertainment, and communication [24].

As shown in Fig. 1, the exterior of modern mobile phone screens is predominantly covered with a glass panel, beneath which are concealed the display panel, polarizer, touch screen, camera, and several optical sensors [28]. The glass panel's surface quality has a direct impact on the display effect, touch effect, imaging quality, and optical sensitivity of the phone's sensors. As shown in Fig. 2, smartphone glass processing involves multiple steps, such as computer numerical control (CNN) cutting, surface finishing, polishing, ink printing, and cleaning. These processes inevitably produce defects such as chippings, scratches, bad-points, dusts, and smudges [5], which can seriously compromise the user experience of smartphones.

Smartphone glass is an ultra-thin rectangular glass, with dimensions of 30-110 mm in width, 50-200 mm in length, and thickness of 0.4−1.0 mm. Much of the light will escape through refraction when ordinary light sources illuminate the glass surface. As a result, collecting defect information from the glass surface of smartphones is becoming increasingly difficult. Furthermore, large smartphone screens significantly

✉ Julong Yuan
jlyuan@zjut.edu.cn

✉ Yongjian Zhu
zhuyongjian_hn@126.com

Yuechu Mao
maoyuechu@126.com

Yingguang Jiang
jiangyg511@163.com

[1] College of Mechanical Engineering, Zhejiang University of Technology, pingfeng, Hangzhou 310023, State, China

[2] Zhejiang University of Technology, Key Laboratory of Special Purpose Equipment and AdvancedProcessing Technology, Ministry of Education and ZhejiangProvince, Ultra-Precision Machining Center, Hangzhou, China

[3] College of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai 200000, China

[4] Ningbo Minjie Information Technology Co., Ningbo 315000, China

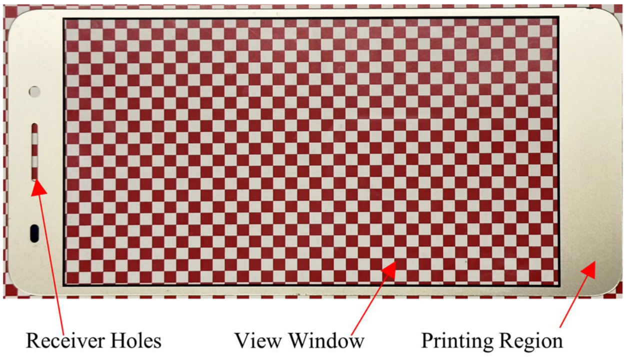


Fig. 1 The surface division of the smartphone glass is mainly composed of the printing region, the view window, and receiver holes

complicate the development of real-time detection systems. Therefore, the development of a real-time detection system with strong anti-interference ability and high accuracy has significant potential for broad practical applications.

Deep learning is mainly divided into single-stage [9–11, 15, 25, 26] and two-stage [18–20] algorithms. These two stages are mainly based on the region-based convolutional neural networks (R-CNN) series, which has high accuracy, but it is difficult to meet the real-time requirements for speed because of the complex detection process. The single-stage algorithm considers accuracy and real-time performance, increasing its use in industrial inspection. Among the single-stage algorithms, the YOLO v5s network has the most outstanding flexibility and detection performance.

Based on YOLO v5s, this study proposes a single-stage improved algorithm called Dy-YOLO v5s to further improve the detection performance. The following improvements are made: (1) To enhance the feature extraction ability of the backbone network, the lightweight attention module pyramid split attention (PSA) [13] and the residual [7] structure are combined to form a CPSA module, which is stacked in a ratio of 1 : 3 : 3 : 1. (2) A new feature pyramid structure called GiraffeDet feature pyramid network (GFPN) [22], which increases cross-scale and cross-layer connections, is introduced to prevent the loss of feature information caused by the deepening of network structure. (3) To improve the detection performance of the head, a dynamic detection framework called dynamic head (DyHead) [4] is added at the end of the detection network to enhance the perception of the spatial position, spatial scale, and task area of the head. (4) The cross-entropy function varifocal loss [29] is used to solve the problem of an imbalance between positive and negative samples when detecting dense targets. (5) For screening anchor boxes, the confidence propagation cluster (cp-cluster) algorithm [21] is used to transform the prediction box deduplication problem into the confidence propagation problem, which improves the confidence and accuracy of the prediction box. Compared with the original YOLO v5s model, the detection algorithm ensures that the P, R, mAP@0.5, and model complexity floating point operations (GFLOPs) are increased by 1.8%, 2.3%, 3.9%, and 8.3, respectively. Compared with other deep learning detection algorithms, the accuracy and speed of detection are better, fully satisfying the business requirements for smartphone glass detection.

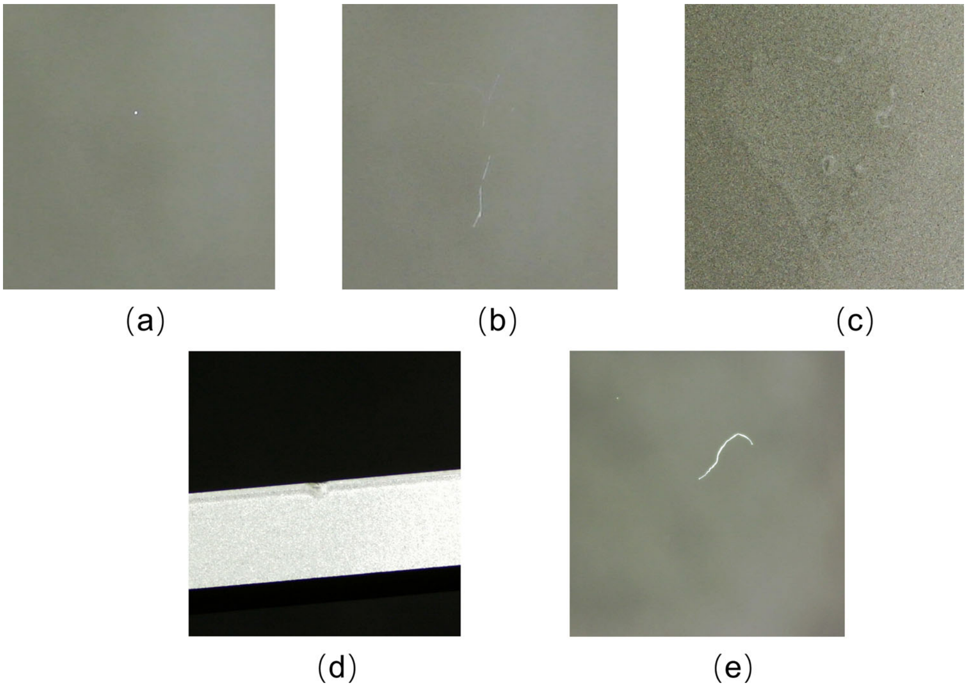


Fig. 2 Under a high-magnification microscope (500x), images of various types of defects on the surface of the glass are captured, including: **a** Bad-Point, **b** Scratch, **c** Smudge, **d** Chipping, **e** Dust

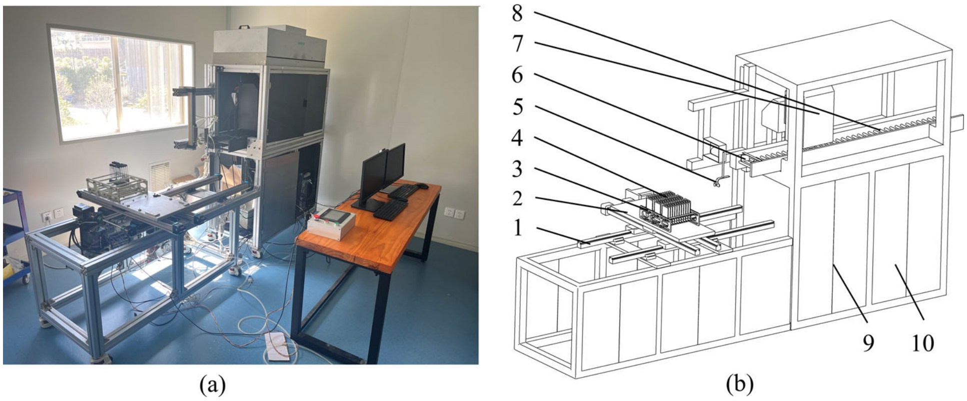


Fig. 3 **a** The physical diagram of the defect detection equipment for glass surfaces of smartphones; **b** The structural diagram of the defect detection equipment for glass surfaces of smartphones. These components include: 1- feeding slide rail; 2-loading platform; 3-placement rack; 4-smartphone glass to be detected; 5-loading adsorption structure; 6-limit device; 7-image acquisition black box; 8-roller conveyor; 9-signal processing box for lower computer; and 10-upper chassis

2 Methodology

2.1 Principle and device of images acquisition

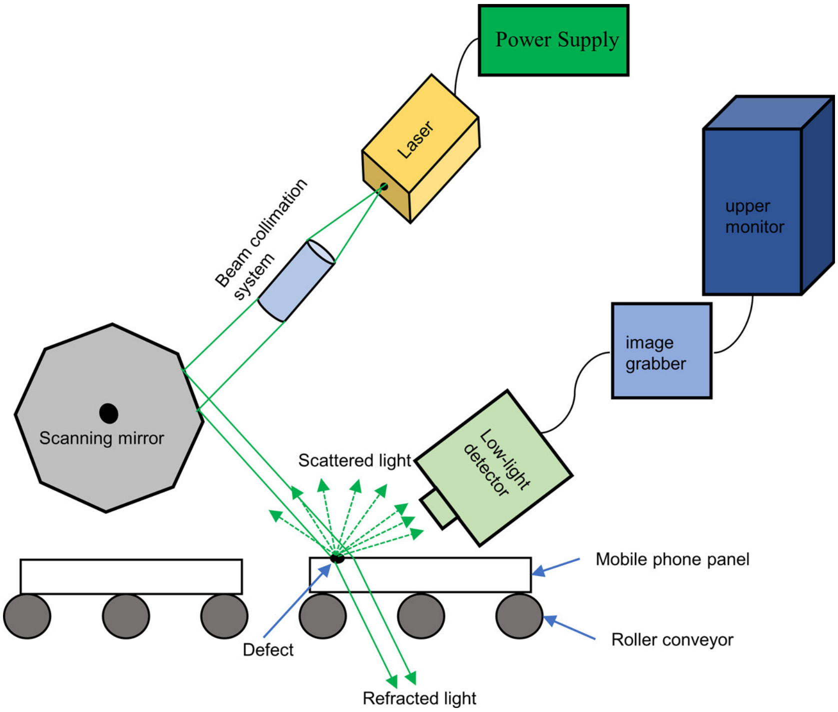
As shown in Fig. 3, to obtain a clearer defect image, this study designs an image acquisition and detection device based on the laser scattering principle [12].

At the beginning of the operation, the intelligent smartphone glass is first placed on the placement rack, before the loading platform moves the glass to the designated position for detection. Then, the loading adsorption structure vacuum adsorbs the product and places it on the roller conveyor mechanism. The limit device at the entrance of the transfer mechanism corrects the scanning posture of the smartphone glass. When the product reaches the acquisition position, the sensor triggers the image acquisition module to collect defect information, which is then transmitted to the upper computer for defect detection and recognition.

The image acquisition module is based on the laser scattering principle, and the specific operating principle is illustrated in Fig. 4. The key optical components used in our experimental arrangement encompass a semiconductor laser diode denoted as MGL-S-532, developed by The Institute of Chang Chun Optical-mechanical. This device emits coherent light at a wavelength of 532 nm with an output power of 300 mW. Furthermore, our experiment used octahedral reflector mirrors obtained from Lincoln Laser Company, labeled as DT-08-039/P1.

During the working process, the smartphone glass is transmitted using the roller-type transmission device, the collimation system forms the laser beam into a parallel beam, and the scanning mirror is rotated to scan the surface of the

Fig. 4 The smartphone glass surface defect detection schematic diagram



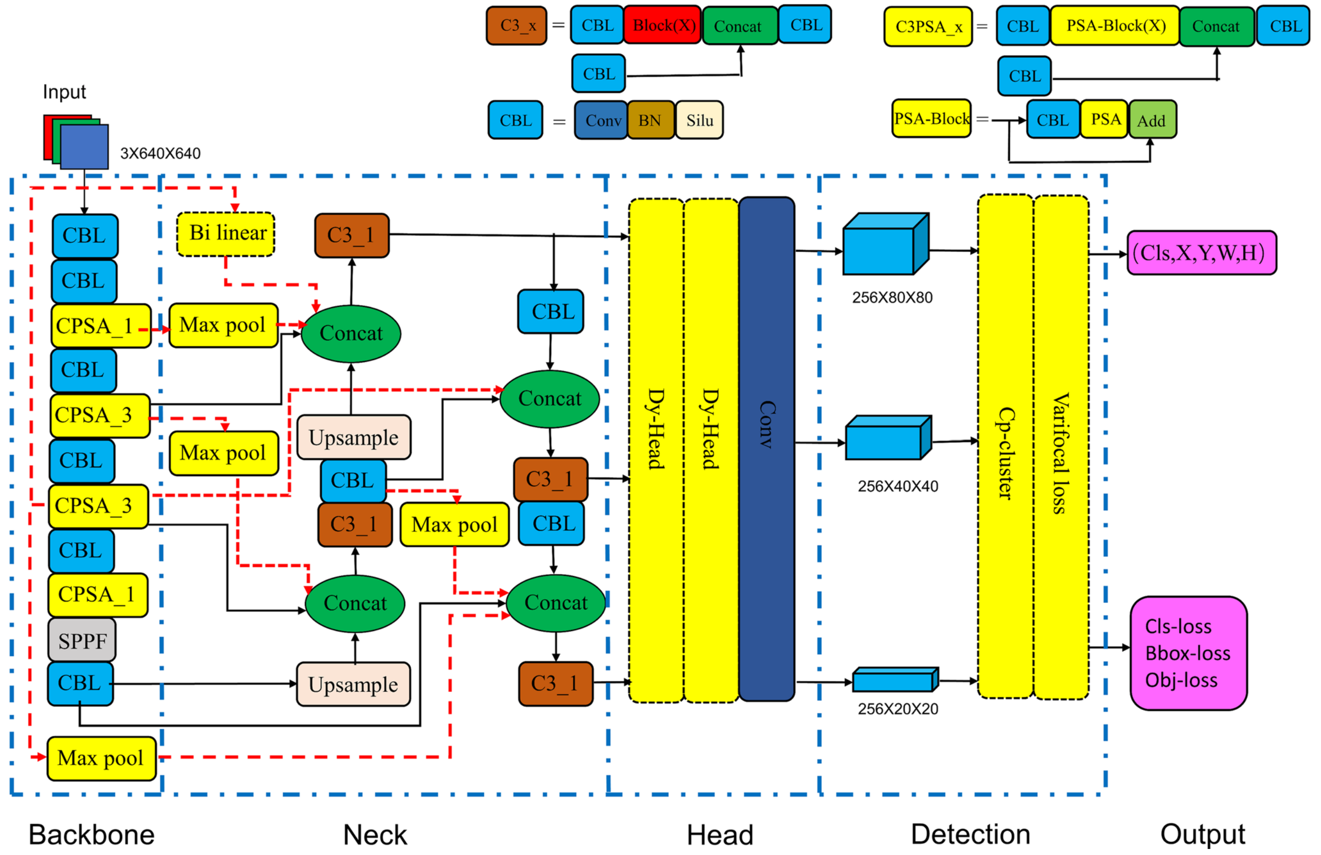


Fig. 5 Dy-YOLO v5s network structure. The improved module is marked in yellow. The red dotted line denotes the new connection between feature maps

smartphone glass line by line. When scanning on a defect-free smartphone glass surface, refracted and reflected light is mainly generated, and when scanning a defect, the scattering phenomenon occurs. Simultaneously, the low-light detector receives the scattered light and transmits the signal to the image acquisition card, which converts it into a digital sig-

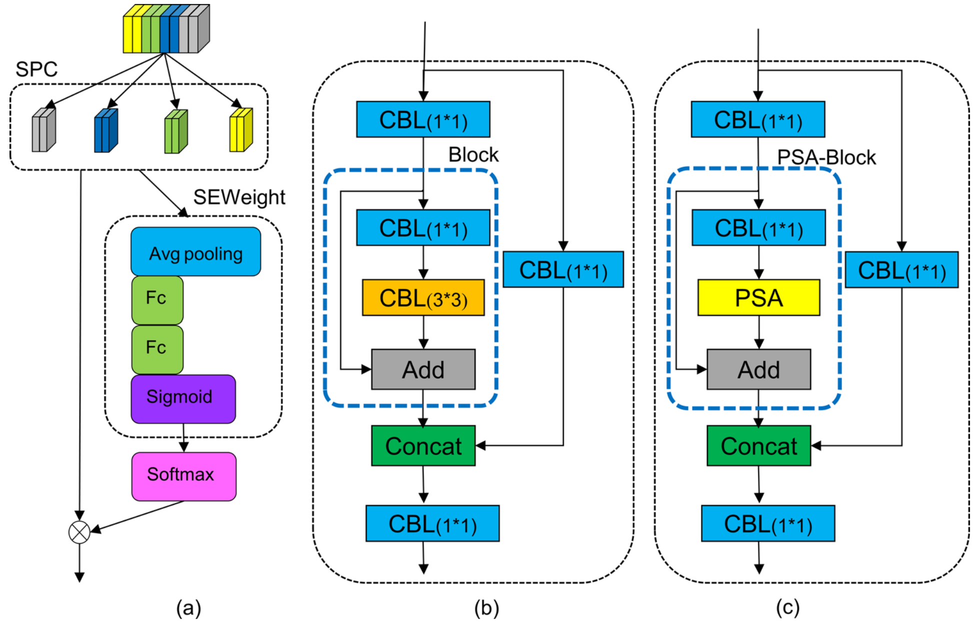


Fig. 6 **a** The PSA module network structure. In the PSA module, feature extraction is initially performed using convolutional kernels of various sizes. Subsequently, the obtained multi-scale feature maps are concatenated using the Concat function. The structures C3_1 and CPSA_1 in the backbone network correspond to **b** and **c**, respectively. By comparing **b** and **c**, the application method of PSA in Dy YOLO v5s can be elucidated

nal and outputs it to the upper machine to display and detect defects.

2.2 Improved Dy-YOLO v5s model

The YOLO v5s is a popular deep-learning detection algorithm in the field of industrial detection owing to its small model size, fast calculation speed, and high recognition rate. YOLO v5s has good detection performance for large-size and sparse targets. However, for detecting smartphone glass, the defect size is small and dense, the overlap is high, and the image background is complex. Therefore, YOLO v5s detection is unsatisfactory. It is necessary to enhance the feature extraction and information exchange capabilities of the network, optimize the anchor boxes' screening method, and reduce the impact of background on target recognition.

The network structure of Dy-YOLO v5s is shown in Fig. 5, which is divided into four parts: backbone, neck, head, and detection. The main innovations are as follows: (1) In the cross stage partial module (CSP) of the backbone network, the PSA module and the residual structure block are combined to form a new CPSA module, which is stacked and used in a ratio of 1 : 3 : 3 : 1. (2) The neck network strengthens the connection between feature maps using the GFPN method and avoids the loss of defect information owing to too deepening of the network. (3) DyHead, a dynamic head detection framework, is added to the head to unify target detection and self-attention. This structure significantly improves the perception and expressiveness of the object detection head. (4), the Cp-cluster algorithm is used during detection to improve the screening accuracy and efficiency of anchor boxes. The algorithm uses the concept of confidence propagation in an undirected graph and transforms the deduplication problem of anchor boxes into a confidence propagation task. Additionally, to alleviate the imbalance between positive and negative samples, the cross-entropy loss function varifocal loss is introduced.

2.3 CPSA—attention module

To improve the feature map extraction capability of the backbone and form a new module CPSA, the attention module PSA is used to replace the 3 × 3 convolution in the residual structure of CSP. As shown in Fig. 6, PSA, as a novel attention module, can efficiently process the spatial data of multi-scale feature maps. It mainly consists of four parts. First, the SPC module integrates the spatial data from different scale feature maps. The specific measures use three, five, seven, and nine convolution kernels for group convolution [27], each convolution group with a size of two, four, eight, and 16, respectively. Second, the SEWeigh [8] module trains the SPC-processed feature map to obtain the weight vector. The Softmax [3] function can normalize the weight vector. Then, the normalized weight vector is dot-multiplied with the feature map output by SPC. The PSA module realizes the interaction between attention weights and spatial channels, producing more informative multi-scale feature maps.

The PSA-Block structure in Fig. 6(c) is the residual structure formed by replacing the attention module PSA. This structure can be stacked in different numbers to form CPSA_1, CPSA_2, and CPSA_3. Figure 7(a) shows the CPSA_1 module in the backbone network, which contains only one PSA-Block structure, while Fig. 7(b) and (c) show CPSA_2 and CPSA_3 modules with two and three stacked PSA-Block structures, respectively. As shown in the backbone network part of Fig. 5, When the stacking ratio of 1 : 3 : 3 : 1 is adopted, the CPSA_1, CPSA_3, CPSA_3, and CPSA_1 in the backbone network of Dy YOLO v5s will be connected in sequence.

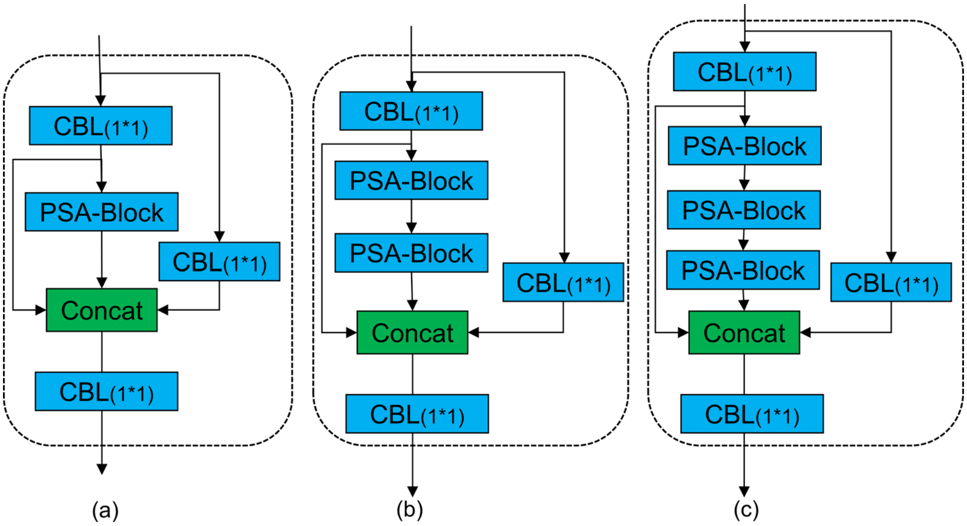


Fig. 7 **a**, **b**, and **c** correspond to CPSA_1, CPSA_2, and CPSA_3 modules respectively, where PSA-Block structures are stacked in different numbers. In the backbone network of Dy-YOLO v5s, CPSA modules are stacked in a ratio of 1 : 3 : 3 : 1, that is, CPSA_1, CPSA_3, CPSA_3, and CPSA_1 are sequentially connected

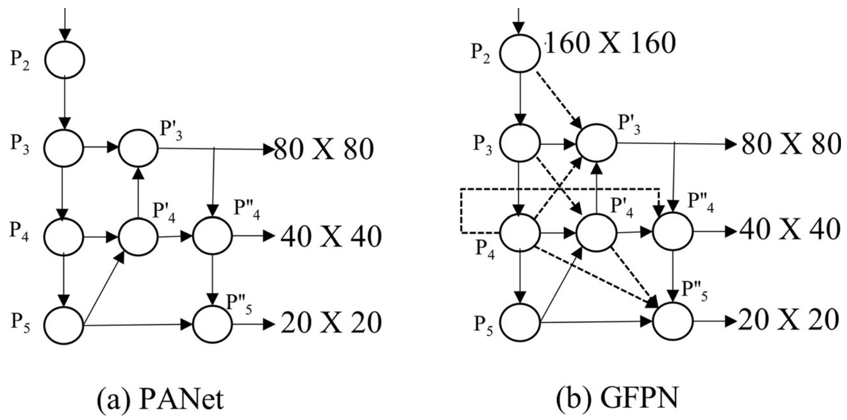


Fig. 8 By comparing **a** PANet and **b** GFPN, GFPN increases the cross-scale connection between adjacent layers and the cross-layer connection at the same scale

2.4 GFPN—a feature pyramid with nore fusions

The detection method used in this study has small-sized defects such as bad-points and chippings, and feature information will continuously be lost when transmitted through the deep network. Therefore, it is necessary to optimize the structure of the neck network to increase the fusion and information exchange between feature maps of various scales and prevent information loss caused by the network being too deep. In the path aggregation network (PANet) [14] that the original YOLO v5s use, the fusion of feature maps only exists between the same scale of the adjacent layer and the adjacent scale of the same layer. Because this fusion route is relatively simple, and the feature information is still lost in the transmission of PANet, this study adds cross-layer connections at the same scale and cross-scale connections between different layers. These feature fusions enable the deep and shallow layers, large- and small-scale information in the network to fully communicate, which improves the accuracy of small-scale target recognition in shallow layers.

As shown in Fig. 8, compared with PANet, the dotted line is the newly added connection of GFPN. The large-size feature map must be down-sampled by max pooling and then fused with the small-size feature map for the cross-scale connection between adjacent layers. The new connections of this type are $\{P_2 - P_3^{'}; P_3 - P_4^{'}; P_4 - P_5^{''}; P_4^{'} - P_5^{''}\}$. The small-size feature map must use bilinear interpolation (Bi linear) as the up-sampling operation before it is fused with the larger-size feature map, which creates a new connection $\{P_4 - P_3^{'}\}$. Scale transformation is not required for the same-scale and cross-layer connection $\{P_4 - P_4^{''}\}$, and the feature maps are fused directly.

2.5 DyHead—Dynamic Detection Framework

In Dy-YOLO v5s, to enhance the head's perception of defects in different morphologies, this study adopts a novel dynamic detection framework, DyHead, which combines the spatial scale, spatial location, and task awareness of objects with a multi-head self-attention mechanism. Compared with the original head, the expression ability and detection accuracy are significantly improved.

To unify scale perception, space perception, and task perception, the DyHead mainly designs three parts of the attention mechanism:

1. Spatial-aware attention: rely on a deformable convolution [30] to extract the target position in the feature map.

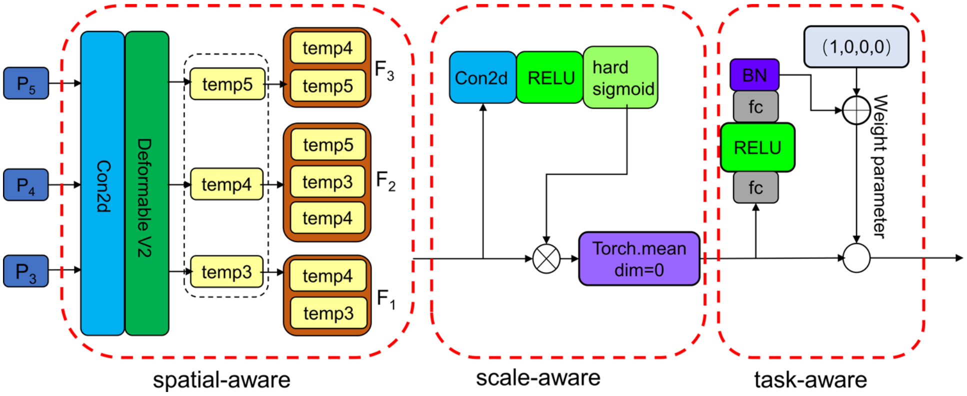


Fig. 9 DyHead detection framework structure diagram, the feature maps must be passed through the three attention modules of spatial perception, scale perception, and task perception in turn

2. Scale-aware attention: through 1×1 convolution, rectified linear unit(ReLU) and hard sigmoid activation function in turn. Through the fusion of feature maps of different scales, the spatial-scale information of the target can be obtained.
3. Task-aware attention: use a fully connected network as a "classifier" to expand the input information for classification.

As shown in Fig. 9, DyHead is used in Dy-YOLO v5s. First, through spatial-aware attention, three different scale feature maps temp3, temp4, and temp5 can be obtained. Then, temp3, temp4, and temp5 are up/down-sampled to a unified scale and fused with each other to form a new three-dimensional tensor F_1, F_2, $F_3 \in R^{L \times S \times C}$, where L denotes the feature level, which indicates the number of feature maps to be fused, S denotes the feature map size, and C denotes the number of channels. Then, the feature tensors F_1, F_2, and F_3 are passed through the scale-aware and task-aware attention mechanisms to generate the information of anchor boxes.

2.6 Cp-cluster—a confidence propagation algorithm

For the screening and deduplication of the anchor boxes, YOLO v5s generally uses non max suppression (NMS) [1] or Soft-NMS [17] methods. In NMS, the anchor boxes must be arranged according to the confidence, and the one with the highest confidence is selected as the ground truth (GT) box. The IOU of the remaining anchor boxes and GT box is calculated, and the anchor boxes are removed above a certain threshold. This method cannot achieve multi-object parallel processing. Furthermore, NMS assumes that the highest confidence score is the GT box, which is not entirely consistent with the actual situation. When dealing with the redundant boxes, NMS is directly set to zero and clear, and the information on these redundant boxes is not fully used. Therefore, the cp-cluster algorithm is introduced into Dy-YOLO v5s to replace NMS or Soft-NMS when screening anchor boxes. The screening efficiency can be improved through parallel computing, and the information on the redundant boxes can be fully used to further improve the confidence of the GT box.

The cp-cluster algorithm must calculate the IOU of each anchor and label box and aggregate the anchor boxes whose IOU is greater than the set threshold β into an undirected graph set. The target of the anchor box detection in this set is consistent. If the IOU of the two anchor boxes in each set is greater than the set threshold θ, a connection is formed between the two nodes. Each anchor box is used as a target anchor box in turn. In this set, the anchor boxes with confidence greater than the target anchor box are called the strong adjacent boxes. In the strong adjacent boxes, if the IOU of the strong adjacent boxes and the target anchor box is greater than the threshold, it will have a negative impact on the confidence of the target anchor box. In the set, if the anchor box confidence is lower than the target anchor box, it is called the weak adjacent anchor box, which has a positive impact on the confidence of the target anchor box. Finally, the confidence of the prediction box is strengthened, and the confidence of the redundant boxes is continuously weakened.

The positive effect on the confidence of the target anchor box, as illustrated in Eq. 1.

$$M_p(i) = Q/(Q+1) \times (1 - \hat{P}(b_i)) \times \max_{b_j \in W_{b_j}} \hat{P}(b_j) \tag{1}$$

In the equation, b_i denotes the object that must be calculated with confidence;b_j and b_i belong to the same set of undirected graphs,$\hat{P}(b_j)$; $\hat{P}(b_i)$ denotes the confidence of the b_j; b_i and $(Q+1)$ denote the number of nodes in the set of undirected graphs; and W_{b_j} denotes the set of nodes of b_j.

The propagation of the negative influence $M_n(i)$ can suppress the redundant boxes.

$$M_n(i) = \hat{P}(b_i) \times IOU\Big(b_i, \operatorname*{arg\,max}_{b_j \in N_{b_i}, SUP_{j,i} \le \xi} M_{b_j}\Big) \tag{2}$$

where $SUP_{j,i}$ denotes a suppression count matrix, which is used to limit the number of times b_i is suppressed. $M_n(i)$ is the negative parameters of the weak adjacent boxes. The maximum negative parameters are selected from many weak adjacent boxes to suppress the confidence of the target anchor box.

The negative parameters are expressed as shown in Eq. 2.

$$M_{b_j} = \Big(\alpha \times \hat{P}(b_i)/\hat{P}(b_j) + (1-\alpha) \times IOU(b_i, b_j)/\theta\Big) \tag{3}$$

Among them, α is a negative influence factor, and the value of α determines the selection of different strong neighbor anchor boxes b_j, which weakens the confidence of b_i. During training, the box with the largest confidence value ($\alpha = 1.0$) and the box with the highest overlap ($\alpha = 0.0$) are alternately selected.

Based on the above, combining the positive and negative effects of the adjacent anchor boxes results in the final confidence of the target anchor box of $\hat{P}(b_i)$.

$$\hat{P}(b_i) = \hat{P}(b_i) + M_p(i) - M_n(i) \tag{4}$$

2.7 Cross-entropy function—varifocal loss

To detect dense small-size objects, this study adopts a new cross-entropy function, varifocal loss. This function uses the focal loss [23] weighting concept to optimize the imbalance problem of positive and negative samples. According to Eq. 5,

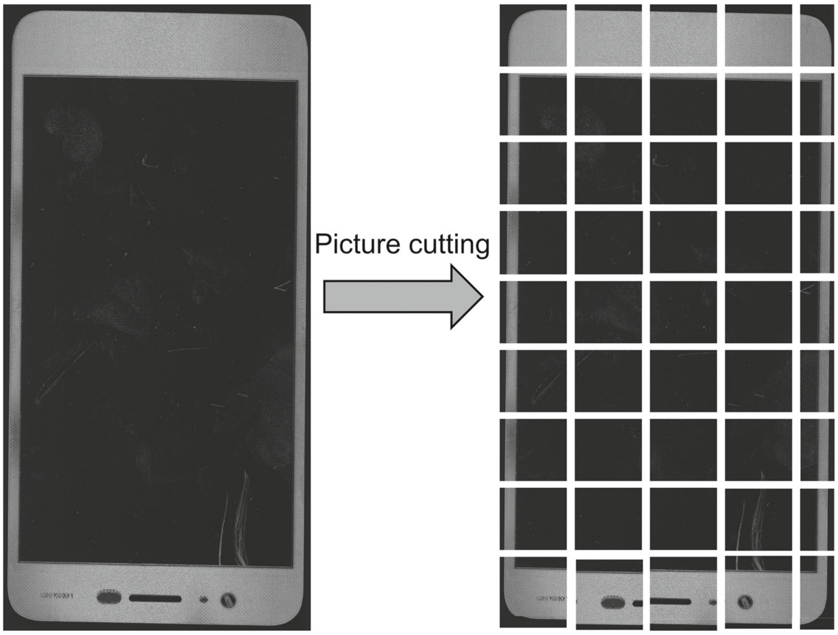


Fig. 10 Dividing an image with a resolution of 4500 × 6600 into multiple 640 × 640 images and sending them to Dy-YOLO v5s can reduce the loss of information caused by adaptive scaling of the image

the contribution of negative samples to the loss is reduced by different weighting amounts, enhancing the contribution of positive samples to the loss and balancing the positive and negative samples. Among them, α denotes the negative loss calculation factor, which is 0.75 by default, γ denotes the calculation modulation factor, which was set to 1.5 in this study, q denotes the IOU value of the prediction box and the labeling box, and P denotes the score of the prediction box. When $q = 0$, the box is represented as a negative sample, and its contribution to the loss is continuously reduced by P^{γ}, and when $q = 1$, it means that the box is represented as a positive sample. The higher the anchor box score, the greater the contribution to the loss.

$$VEL_{(p,q)} = \begin{cases} -q\big(q\log(P) + (1-q)\log(1-q)\big) & q > 0 \\ -\alpha P^{\gamma}\log(1-P) & q = 0 \end{cases} \tag{5}$$

3 Results and discussion

Experimental environment and data preparation. The experimental configuration of this study is as follows: Windows 10 operating system, 2.3 GHz Intel Xeon Gold 5118 CPU, 52 GB of memory, NVIDIA RTX 2080Ti 12G × 2 graphics processor, and CUDA 1.10.0 + PyTorch 1.9.0 training environment. The model training parameters are as follows: for each training, input images batch size of 64, 300 training epochs, $num_{works} = 6$ training thread, and $lr = 0.001$ learning rate.

The test samples used in this study were all randomly sampled from the production line of the enterprise. The chosen samples were sent to the above device for image acquisition. The size of the image scanned by the device was 4500 × 6600. If the sample is directly sent to the detection algorithm, the size becomes 640 × 640 after the image adaptive operation. Such large-scale image scaling will result in the loss of small-sized defects, which affects the detection accuracy. As shown in Fig. 10, before creating the dataset, the image must be reduced to 640 × 640.

The divided images were screened and labeled to create a dataset. The dataset used in the experiment comprises 16,410 images, including 13,128 for the training set and 1,641 images each for the validation and test sets. Each category of defects in the training set contains at least 2,000 images, allowing the detection model to learn sufficiently and improve its generalization and accuracy. As shown in

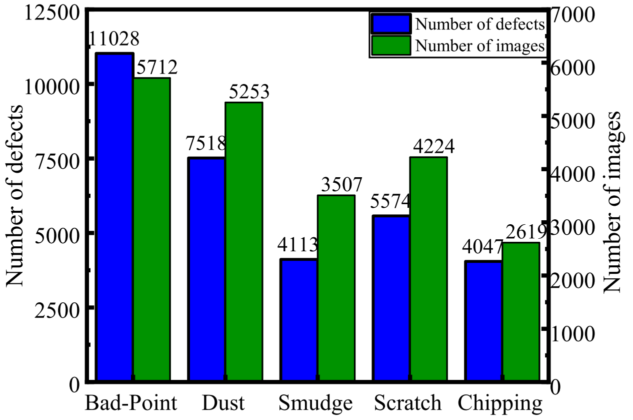


Fig. 11 The statistical chart of various types of defects in the smartphone glass dataset

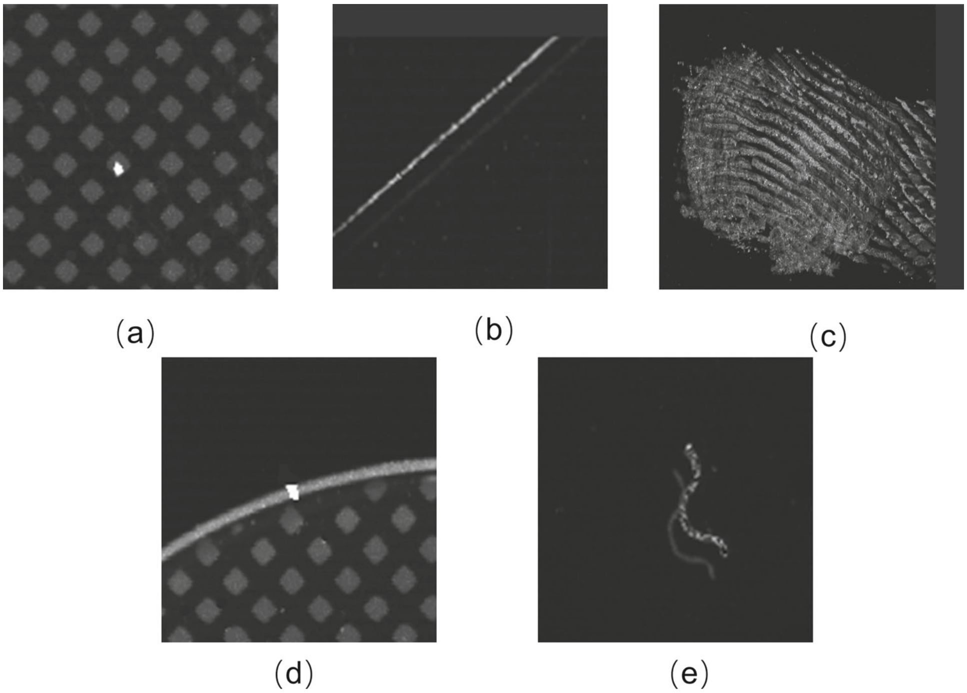


Fig. 12 The collected images of surface defects of different types of smartphone glass: **a** Bad-Point, **b** Scratch, **c** Smudge, **d** Chipping, and **e** Dust

Fig. 11, the bar chart illustrates the number of each type of defect and the corresponding number of images in which they appear. The chart demonstrates that bad-point defects are many and have the widest distribution, whereas chipping defects are few and have the least distribution. However, both the quantity and distribution ratios of each type of defect are greater than 1:4, implying that the collected dataset has a balanced distribution and is unlikely to generate overfitting.

From the defect sample in Fig. 12, the defects of the smartphone glass have the following characteristics:

1. The defect size gap is large, and the smallest defect in the bad point is approximately 3 × 3 pixels, while the largest one in the smudge is 400 × 400 pixels.
2. The width of the scratch is narrower by approximately 2 pixels.
3. Chipping is located at the edge of the panel, and the grayscale value of the defect is almost the same as that of the edge pixel, which is difficult to detect.
4. Smudge defects in various shapes and sizes.

Backbone network improvement. This experiment aims to verify that the improved backbone of Dy-YOLO v5s has a stronger feature extraction ability than Darknet53 (backbone of YOLO v5s). Based on YOLO v5s, different backbones were constructed and tested by adding attention modules and changing the stacking ratio of residual modules. The main module of the Darknet53-1 network is CSP with a stacking ratio of $1 : 2 : 3 : 1$, and the ratios of Darknet53-2 and Darknet53-3 are $1 : 1 : 3 : 1$ and $1 : 3 : 3 : 1$, respectively. Compared with Darknet53, PSA- Darknet53 replaces CSP with CPSA.

As shown in Table 1, as the attention mechanism PSA is incorporated into Darknet53, the mAP increases by approximately 2% when the IOU threshold is 50%, and the average mAP between the IOU thresholds between 50% and 95% increases by approximately 1.5%. Although the model complexity GFLOPs increased slightly, the detection speed FPS

Table 1 Backbone network ablation experiment

Backbone	Radio	mAP(%) @0.5	mAP(%) @0.50 – 0.95	GFLOPs (10^9)	FPS
Darknet53-1	1:2:3:1	91.7	49.7	15.8	61.73
Darknet53-2	1:1:3:1	91.0	49.1	15.3	65.36
Darknet53-3	**1:3:3:1**	92.6	51.3	16.4	59.50
PSA- Darknet53-1	1:2:3:1	93.7	52.3	16.0	63.69
PSA- Darknet53-2	1:1:3:1	93.2	51.5	15.4	65.36
PSA- Darknet53-3	**1:3:3:1**	**94.2**	**52.6**	16.5	61.73

Table 2 Comparative experiments of Neck network

Backbone	Neck	mAP(%) @0.5	mAP(%) @0.50 – 0.95	GFLOPs (10^9)	FPS
Darknet53	PANet	91.7	49.7	15.8	61.73
	GFPN	94.0	52.3	20.3	56.50
PSA- Darknet53-3	PANet	94.2	52.6	16.5	61.73
	GFPN	**94.8**	**53.2**	20.8	56.50

remained unchanged. This shows that the PSA module improves the feature extraction capability of the backbone network. According to the comparative experiments of radio, increasing the proportion of modules used helps the extraction of target features. Comprehensive experimental data demonstrate that the final improved PAS-Darknet53-3 compared with Darkenet53 improved mAP and enhanced the network performance, while the detection speed remained unchanged.

Neck network improvement. The feature pyramid network PANet used in the neck stage of YOLO v5s has a relatively simple fusion line of feature maps. Therefore, Dy-YOLO v5s introduced more GFPN fusion routes. Comparative experiments are required to study the effect of GFPN on the information extraction and communication fusion capabilities of detected targets in feature maps. According to the experiments in this section, the backbone uses Darknet53 and the improved network PSA-Darknet53-3, while the neck uses PANet and GFPN for combined experiments. Finally, the performance of different neck networks is compared according to the experimental results.

According to the experimental data in Table 2, compared with the original PANet, the GFPN network is more frequently fused between feature maps of different scales and levels. This increases the FLOPs of model complexity by approximately 4.5G and decreases the FPS by 5.23, but the output feature map is more informative and accurate. Based on Darknet53, when only GFPN is used, the performance is improved significantly; mAP@0.5 and mAP@0.5−0.95 increased by 2.3% and 2.6%, respectively. When based on the improved backbone, both mAP@0.5 and mAP@0.5−0.95 increased by 0.6%. Overall, for images of smartphone glass, GFPN loses less defect information than PANet.

Head improvement. The experiments in this section used the models Darknet53-1+ PANet and the optimized network PSA- Darknet53-3+ GFPN, while the heads with and without DyHead were combined to form four groups of experiments.

According to the experimental data in Table 3, regardless of whether the DyHead framework is added to the original Darknet53-1+ PANet or the improved PSA-Darknet53-3+ GFPN model is used, the evaluation index mAP of the detection results is significantly improved. Using DyHead alone increases mAP@0.5 by 3.1%, while adding DyHead based on the improved algorithm increases mAP@0.5 by 0.4%. Although part of the detection speed is sacrificed, adding the DyHead module to the YOLO series algorithm significantly improves the detection of smartphone glass.

Detection function improvement. In the detection of the algorithm, three common operations such as NMS, Soft-NMS, and cp-cluster are used for the deduplication and screening of anchor boxes. For the balance of positive and negative samples, two cross-entropy functions: focal loss and varifocal loss are used for experiments. The model of the algorithm uses the network optimized by the above experiments.

According to the experimental data in Table 4, among the three deduplication methods- NMS, Soft-NMS, and cp-cluster- cp-cluster demonstrated a significant improvement in the detection speed. This shows that the parallel operation method of the cp-cluster is suitable for smartphone glass detection. According to the comparison experiment of the cross-entropy function, analyzing that the varifocal loss significantly improves the accuracy indicators such as mAP is not difficult. Therefore, the cp-cluster algorithm significantly improves the detection speed through the parallel screening of anchor boxes. Through the evolution of the weight factor, the varifocal loss function can effectively enhance the contribution of anchor boxes with higher loss confidence, while suppressing the contribution of negative samples or low-confidence anchor boxes to the loss.

Table 3 Improvement Experiment of Head Structure

Backbone	Head	mAP(%) @0.5	mAP(%) @0.50 – 0.95	GFLOPs (10^9)	FPS
Darknet53-1+ PANet	N/A	91.7	49.7	15.8	61.73
	DyHead	94.8	52.3	19.1	37.47
PSA- Darknet53-3+ GFPN	N/A	94.8	53.2	20.8	56.50
	DyHead	**95.2**	**53.8**	24.2	32.33

Table 4 Optimization of anchor screening algorithm and cross-entropy function

Cross-entropy	Anchor Box	mAP(%) @0.5	mAP(%) @0.50 – 0.95	GFLOPs (10^9)	FPS
Focal loss	NMS	95.2	53.8	24.3	32.33
	Soft-NMS	91.6	51.2		18.87
	Cp-cluster	**95.5**	54.2		43.29
Varifocal loss	NMS	95.6	53.8		22.37
	Soft-NMS	92.6	52.8		16.08
	Cp-cluster	**96.2**	55.3		30.58

Dy-YOLO v5s. Based on the aforementioned improvement experiments, the improvement steps of Dy-YOLO v5s and the change curves of some detection indexes are listed in Fig. 13. The improvement steps based on YOLO v5s are represented from left to right in the diagram. First, the stacking ratio of CSP modules is adjusted to 1: 3: 3: 1, which increases the mAP@0.5 and model complexity GFLOPs by 0.4% and 0.6, respectively. Then, the C3PSA module replaces the CSP module, which increases mAP@0.5 and model complexity GFLOPs by 1.0% and 0.7, respectively. Then, GFPN and DyHead networks are added to the neck and head to increase mAP@0.5 and GFLOPs to 95.2% and 24.1, respectively. Finally, the varifocal loss function and cp-cluster algorithm are used to improve the detection accuracy to 96.2%, while maintaining the model size.

In this experiment, different YOLO algorithms, the two-stage algorithm faster RCNN, Mobilenet [2, 6, 16], and Dy-YOLO v5s were compared to show the detection performance of the Dy-YOLO v5s algorithm. According to the experimental results in Table 5, the Dy-YOLO v5s algorithm has better accuracy in the single-stage detection algorithm, and the model complexity and detection speed are moderated, which can satisfy the real-time requirements. Compared with the two-stage classic algorithm faster RCNN, although the accuracy is slightly lower by 0.6%, it has significant advantages in model complexity and detection speed.

According to the above experiments, the improved Dy-YOLO v5s in this study has obvious advantages in the detection of smartphone glass defects. According to Table 6, the detection results of P, R, and mAP for all defects are listed in detail when training with Dy-YOLO v5s. Additionally, Fig. 14 shows the detection effect of all defects. According to the above experimental results, the mAP@0.5 of the Dy-YOLO v5s network was 96.2%; P was 92.6%; the best performance for smudge defect detection reached 98.6%, and the average accuracy rate for scratch and chipping were poor with values 95.8% and 93.4%, respectively.

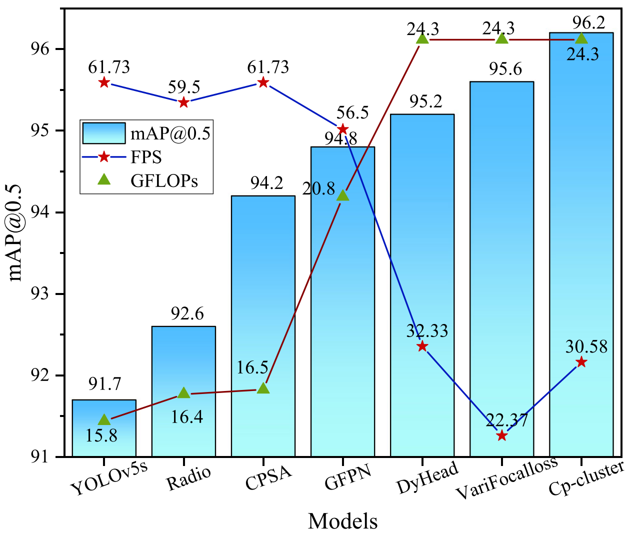


Fig. 13 Changes in various indicators during the improvement of the Dy-YOLO v5s algorithm

Table 5 Comparison of experimental results of different algorithms in smartphone glass defect detection

Models	mAP(%) @0.5	GFLOPs (10^9)	FPS
YOLO v5s	91.7	15.8	61.73
YOLO v3	89.69	116.9	15.64
YOLO v7-tiny	88.6	5.8	132.53
Faster R-CNN	**96.8**	214.6	12.32
Mobilenet-YOLO v5s	86.1	6.3	69.44
Dy-YOLO v5S	**96.2**	**24.3**	**30.58**

Table 6 Detection indexes of various defects in the Dy-YOLO v5s algorithm

Detection	P (%)	R (%)	mAP(%) @0.5
All	**92.6**	**93.1**	**96.2**
bad-point	94.1	92.8	96.4
dust	94.8	92.6	**96.8**
scratch	89.2	92.9	95.8
chipping v5s	88.5	90.6	93.4
smudge	96.2	96.8	**98.6**

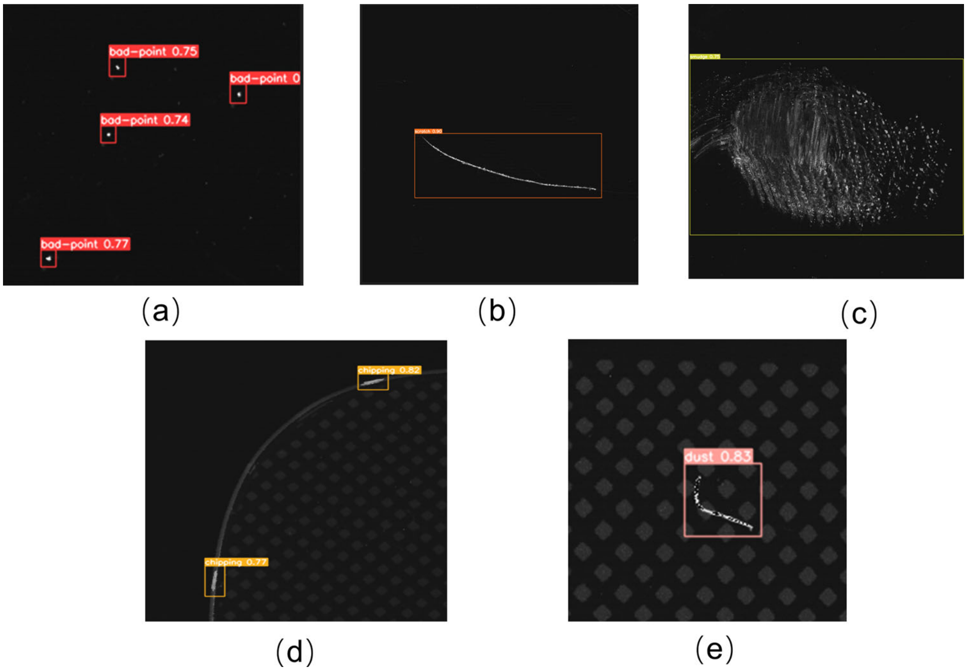


Fig. 14 The different types of smartphone glass surface defect detection renderings: **a** Bad-Point, **b** Scratch, **c** Smudge, **d** Chipping, and **e** Dust

4 Conclusion

In this study, an improved algorithm, Dy-YOLO v5s, is proposed for detecting surface defects on smartphone glass. Among them, to improve the ability of the backbone network to extract input features, the attention mechanism PSA and the residual structure are combined, and the stacking ratio is adjusted in the backbone network. To address the information loss problem, which is caused by small-sized defects owing to network depth, GFPN is used in the neck to strengthen the fusion between feature maps of different scales and layers. To enhance the head's ability to perceive the spatial position, scale, and detection task of the target, a dynamic detection framework DyHead is incorporated. In algorithm detection, the cross-entropy function varifocal loss and the anchor box deduplication method cp-cluster are combined to improve the use efficiency of positive and negative sample information and efficiently and accurately screen out the prediction box. Finally, the map@0.5, GFLOPs, and FPS of Dy-YOLO v5s are 96.2%, 24.3, and 30.58, respectively.

In this study, experiments were conducted using the smartphone glass image dataset to verify the effectiveness of the above-improved methods, and the Dy-YOLO v5s improvement steps are listed. Compared with other classic algorithms such as the YOLO algorithm and faster RCNN, the Dy-YOLO v5s satisfies the detection requirements of smartphone glass in terms of detection accuracy and real-time performance.

Acknowledgements The author would like to thank Julong Yuan and Yongjian Zhu for their helpful discussions on topics relevant to this work. This work was supported by the scientific research start-up project of the Shanghai Institute of Technology under Grant YJ2022-40; the collaborative innovation fund of the Shanghai Institute of Technology under Grant XTCX2022-20. We would like to thank Editage (www.editage.cn) for English language editing.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Yuechu Mao, Julong Yuan, Yongjian Zhu and Yingguang Jiang. The first draft of the manuscript was written by Yuechu Mao and all authors commented on previous versions of the manuscript. All authors have read and approved the final manuscript.

Funding This work was supported by the scientific research start-up project of the Shanghai Institute of Technology under Grant YJ2022-40; the collaborative innovation fund of the Shanghai Institute of Technology under Grant XTCX2022-20.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Alexander N, Luc VG (2006) Efficient non-maximum suppression. Paper presented at 2006 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 850-855 Aug 20-24 2006, 10.1109/ICPR.2006.479
2. Andrew H, Mark S, Grace C, Chen LC, Chen B, Tan MX, Wang WJ, Zhu YK, Pang RM, Vijay V (2019) Searching for mobilenetv3. Paper presented at 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 1314-1324 Oct 27 2019, 10.1109/ICCV.2019.00140
3. Armand J, Moustapha C, David G, Hervé J (2017) Efficient softmax approximation for gpus. Paper presented at Proceedings of the

34th International Conference on Machine Learning - Volume 70, Sydney, NSW, Australia, 1302–1310 2017, http://proceedings.mlr.press/v70/grave17a.html?ref=https://githubhelp.com

4. Dai XY, Chen YP, Xiao B, Chen DD, Liu MC, Yuan L, Zhang L (2021) Dynamic head: Unifying object detection heads with attentions. Paper presented at 2021 IEEE/CVF conference on computer vision and pattern recognition, 7369-7378 June 20-25 2021, 10.1109/cvpr46437.2021.00729
5. Di L, Quan LL, Jie ZW (2014) Defect inspection and extraction of the mobile phone cover glass based on the principal components analysis. The International Journal of Advanced Manufacturing Technology 73:1605–1614. https://doi.org/10.1007/s00170-014-5871-y
6. G HA, Zhu ML, Chen B, Dmitry K, Wang WJ, Tobias W, Marco A, Hartwig A (2020) Mobilenets: Efficient convolutional neural networks for mobile vision applications. International Journal on Advanced Science, Engineering and Information Technology 10:2290–2296, 10.18517/ijaseit.10.6.10948
7. He KM, Zhang XY, Ren SQ, Sun J (2016) Deep residual learning for image recognition. Paper presented at 2016 IEEE Conference on Computer Vision and Pattern Recognition, 770-778 June 27-30 2016, 10.1109/CVPR.2016.90
8. Hu J, Shen L, Sun G (2019) Squeeze-and-excitation networks. Paper presented at 2020 IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011-2023 Aug 1 2020, 10.1109/TPAMI.2019.2913372
9. Joseph R, Ali F (2017) Yolo 9000: better, faster, stronger. Paper presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6517-6525 July 21-26 2017, 10.1109/CVPR.2017.690
10. Joseph R, Ali F (2018) Yolov3: An incremental improvement. arXiv p 1804.02767, 10.48550/arXiv.1804.02767
11. Joseph R, Santosh D, Ross G, Ali F (2016) You only look once: Unified, real-time object detection. Paper presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779-788 June 27-30, 10.1109/cvpr.2016.91
12. Katsumi T (1997) Defect inspection of wafers by laser scattering. Materials Science and Engineering: B 44:181–187. https://doi.org/10.1016/s0921-5107(96)01745-x
13. andKe Ke Zu HZ, Lu J, Zou YR, Meng DY (2021) Epsanet: An efficient pyramid squeeze attention block on convolutional neural network. Paper presented at 2021 IEEE Conference on Computer Vision and Pattern Recognition, 2021, 10.48550/arXiv.2105.14447
14. Liu S, Qi L, Qin HF, Shi JP, Jia JY (2018) Path aggregation network for instance segmentation. Paper presented at 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8759-8768 June 18-23 2018, 10.1109/cvpr.2018.00913
15. Liu W, Dragomir A, Dumitru E, Christian S, Scott R, Cheng-Yang F, C BA (2016) Ssd: Single shot multibox detector. Paper presented at ECCV 2016: Computer Vision, 21-24 September 17 2016, 10.1007/978-3-319-46448-0_2
16. Mark S, Andrew H, Zhu ML, Andrey Z, Chen LC (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. Paper presented at 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4510-4520 Oct 18-23 2018, 10.1109/CVPR.2018.00474
17. Navaneeth B, Bharat S, Rama C, S DL (2017) Soft-nms - improving object detection with one line of code. Paper presented at 2017 IEEE International Conference on Computer Vision (ICCV), 5562-5570 Oct 22-29 2017, 10.1109/iccv.2017.593
18. Ren SQ, He KM, Ross G, Sun J (2017) Faster r-cnn: Towards real-time object detection with region proposal networks. Paper presented at 2017 IEEE Transactions on Pattern Analysis and Machine Intelligence, 1137-1149 June 1 2017, 10.1109/TPAMI.2016.2577031
19. Ross G (2015) Fast r-cnn. Paper presented at 2015 IEEE International Conference on Computer Vision (ICCV), 1440-1448 Dec 7-13 2015, 10.1109/ICCV.2015.169
20. Ross G, Jeff D, Trevor D, Jitendra M (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. Paper presented at 2014 IEEE Conference on Computer Vision and Pattern Recognition, 580-587 June 23-28 2014, 10.1109/CVPR.2014.81
21. Shen YC, Jiang WL, Xu Z, Li RD, Junghyun K, Li SY (2022) Confidence propagation cluster: Unleash full potential of object detectors. Paper presented at 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1141-1151 June 18-24 2022, 10.1109/CVPR52688.2022.00122
22. Tan ZY, Wang JY, Sun XY, Lin M, Li H (2022) Giraffedet: A heavy-neck paradigm for object detection. Paper presented at 2022 International Conference on Learning Representations, 110.48550/arXiv.2202.04256
23. Tsung-Yi L, Priya G, Ross G, Kaiming H, Piotr D (2017) Focal loss for dense object detection. Paper presented at 2017 IEEE International Conference on Computer Vision (ICCV), 2999-3007 Oct 22-29 2017, 10.1109/ICCV.2017.324
24. Wang CY, Lu Q, Tao Y (2020) Five trends in the development of glass for mobile phones. Glass 47(4):1–6, preprint at http://www.cqvip.com/qk/90499x/202004/7101555265.html
25. Wang CY, Alexey B, Mark LHY (2021) Scaled-yolov4: Scaling cross stage partial network. Paper presented at 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13024-13033 June 20-25 2021, 10.1109/CVPR46437.2021.01283
26. Wang CY, Alexey B, Mark LHY (2022) Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv p 2207.02696, 10.48550/arXiv.2207.02696
27. Yani I, Duncan R, Roberto C, Antonio C (2017) Deep roots: Improving cnn efficiency with hierarchical filter groups. Paper presented at 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5977-5986 June 21-26 2017, 10.1109/CVPR.2017.633
28. Yuan ZC, Zhang ZT, Su H, Zhang L, Shen F, Zhang F (2018) Vision-based defect detection for mobile phone cover glass using deep neural networks. International Journal of Precision Engineering and Manufacturing 19:801–810. https://doi.org/10.1007/s12541-018-0096-x
29. Zhang HY, Wang Y, Feras D, Niko S (2021) Varifocalnet: An iou-aware dense object detector. Paper presented at 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8510-8519 June 19-25 2021, 10.1109/cvpr46437.2021.00841
30. Zhu XZ, Hu H, Stephen L, Dai JF (2019) Deformable convnets v2: More deformable, better results. Paper presented at 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9300-9308 June 16-20 2019, 10.1109/cvpr.2019.00953

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