Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8847651, 14 pages
https://doi.org/10.1155/2020/8847651

*Research Article*

# SK-FMYOLOV3: A Novel Detection Method for Urine Test Strips

**Rui Yang,[1] Yonglin Zhang,[1] Zhenrong Deng,[1] Wenming Huang,[1] Rushi Lan,[2] and Xiaonan Luo[3]**

[1]*Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, China*
[2]*School of Computer Science Engineering, South China University of Technology, Guangzhou 510006, China*
[3]*National and Local Joint Engineering Research Center of Satellite Navigation and Location Service, Guilin University of Electronic Technology, Guilin 541004, China*

Correspondence should be addressed to Zhenrong Deng; 799349175@qq.com

Received 21 August 2020; Revised 26 October 2020; Accepted 4 November 2020; Published 18 December 2020

Academic Editor: Zhili Zhou



To accurately detect small defects in urine test strips, the SK-FMYOLOV3 defect detection algorithm is proposed. First, the prediction box clustering algorithm of YOLOV3 is improved. The fuzzy C-means clustering algorithm is used to generate the initial clustering centers, and then, the clustering center is passed to the K-means algorithm to cluster the prediction boxes. To better detect smaller defects, the YOLOV3 feature map fusion is increased from the original three-scale prediction to a four-scale prediction. At the same time, 23 convolutional layers of size $3 \times 3$ in the YOLOV3 network are replaced with SkNet structures, so that different feature maps can independently select different convolution kernels for training, improving the accuracy of defect classification. We collected and enhanced urine test strip images in industrial production and labeled the small defects in the images. A total of 11634 image sets were used for training and testing. The experimental results show that the algorithm can obtain an anchor frame with an average cross ratio of 86.57, while the accuracy rate and recall rate of nonconforming products are 96.8 and 94.5, respectively. The algorithm can also accurately identify the category of defects in nonconforming products.

## 1. Introduction

Defect recognition is one of the important applications of machine vision in the field of industrial manufacturing. It can improve factory production efficiency and reduce human labor and can be used to monitor product quality in real time [1]. However, the accurate identification of product defects is still a challenging problem that is under investigation in current research. To date, two main approaches have been used in research studies, namely, the use of traditional image recognition methods to extract and classify image features, and the direct use of deep neural networks for defect identification.

The traditional defect recognition algorithm includes the following steps: image preprocessing, image segmentation, feature extraction, and classifier training. The goal of image preprocessing is to reduce the noise contained in the images collected in the industrial field [2, 3]. Image segmentation is carried out in order to decompose the image into several areas with different characteristics, with the same or similar image characteristics in each area. Commonly used methods in image segmentation include threshold-based segmentation [4, 5] and edge-based segmentation [6, 7]. Commonly used edge detection operators include Canny operator, Sobel operator, and Roberts operator. Image feature extraction is performed to map a high-dimensional image space to a low-dimensional feature space. Commonly used image features include color, shape, and texture [8, 9]. Color-based feature extraction methods include color histograms [10], color moments, and color sets. Texture-based feature extraction methods mainly include statistical methods [11], frequency spectrum methods [12–14], and model methods [15]. Commonly spectral methods include Fourier transform, wavelet transform, and Gabor transform. Image

texture features can also be extracted from the second-order moments in the gray histogram, entropy, inverse moments, contrast, and correlation [16]. Huang et al. proposed a CDD-based defect detection algorithm to detect and classify defects [17]. The classifier learns the mapping relationship between the feature vector and the category through the training of features and labels [18–20] and finds the model parameters with the smallest classification error. Commonly used classifiers include the ANN, Bayes, and SVM classifiers.

Traditional image recognition methods are inefficient and inaccurate. To improve the recognition algorithm, Girshick et al. [21] proposed the R-CNN algorithm that for the first time introduced deep learning into the field of computer vision. Then, He et al. [22] proposed the SPP-Net algorithm that solved the problem of object deformation caused by the candidate frame scaling to a uniform size. Girshick [23] proposed fast R-CNN by further improving the shortcomings of R-CNN and SPP-Net. Ren et al. [24] proposed faster R-CNN that improves the detection speed while ensuring a certain accuracy. The subsequently developed R-FCN [25] method is also a region-based target recognition method.

Later, regression-based target recognition methods, such as SSD [26] and YOLO [27] series appeared. The region-based method has higher positioning accuracy but has the disadvantage of low detection speed. The YOLO network based on regression has fast processing speed and high accuracy [28], is easy to deploy in industrial production, and has been widely used.

For the recognition of small targets, shortcomings such as small field of vision, single aspect ratio, and low detection accuracy are always present [29]. To solve these problems, many researchers have improved the network performance by improving the structure, introduced the top-down structure, and proposed algorithms such as DSSD [30, 31] and YOLOV3 [32] to improve performance. For example, Tao et al. [33] developed the OYOLO network by increasing the weight of the positioning error function. After combining with R-FCN, the detection speed is improved, but the weight of the confidence error function is reduced, affecting the confidence prediction of the network. Deng et al. [34] proposed a small target recognition algorithm based on CGAN that has performs accurate target recognition but only in a single application scenario. Zheng et al. [35] proposed a dense-YOLO network that improves the recognition of small targets in remote sensing images through feature reuse but has the disadvantage of a huge memory footprint.

To meet the needs of real-time detection in industrial production, this paper improves the detection accuracy of small defects while ensuring fast detection speed. The YOLOV3 network is used as the base detection model, because the YOLOV3 network performs the detection and classification of images simultaneously, greatly improving the detection speed. First, the prediction layer clustering algorithm of YOLOV3 is improved to avoid the influence of the randomly initialized prediction box on the prediction result and improve the accuracy of the prediction box. Additionally, for the smaller defects, YOLOV3 shows missed detections. Therefore, this paper adds a scale to the original YOLOV3, uses 4 scales to detect the target image, and improves the recall rate of small defects. Finally, aiming at the precision of small defects, SKNet structure is added on the basis of YOLOV3 to improve the score of small defects and obtain higher recognition accuracy. For the identification of qualified products and nonconforming products, the precision rate is 96.8, and the recall rate is 94.5. Moreover, our method can accurately identify six minor defects in the nonconforming products, including "Crooked," "Stains," "Marker pen," "Burr," "Short," and "Peeling." The contributions of this paper are as follows:

(i) A prediction box clustering method using a combination of fuzzy C-means and K-means is proposed

(ii) A fusion framework for small target recognition is proposed, and the SkNet structure is embedded in the YOLOV3 network model

(iii) A urine test strip image data set with a size of 11634 was collected that provided data basis for future research and demonstrated new approaches for the identification of small defects in industrial products

Article structure: the article is divided into five parts. The second part introduces the YOLOV3 algorithm and SeNet structure. The third part introduces the designed SK-FMYOLOV3 network model. The fourth part analyzes the performance of the SK-FMYOLOV3 network model and compares and displays the experimental results of industrial product defect detection. The fifth part summarizes the algorithm.

## 2. Propaedeutics

*2.1. YOLOV3.* YOLOV3 is a new end-to-end target detection model after R-CNN, fast R-CNN, and faster R-CNN, as shown in Figure 1. It combines training with target classification and detection and returns the position and category of the target detection box directly at the output layer, transforming the detection problem into a regression problem.

YOLOV3 will predict 4 values for each border on each cell, that is, the coordinates of the upper left corner of the border $(x, y)$ and the width and height of the target $(w, h)$, recorded as $(t_x, t_y, t_w, t_h)$. If the center of the target is offset $(C_x, C_y)$ from the upper left corner of the image in the cell, and the anchor box has a width and height $(P_w, P_h)$, the revised border is

$$\begin{aligned} b_x &= \sigma(t_x) + C_x, \\ b_y &= \sigma(t_y) + C_y, \\ b_w &= P_w * e^{t_w}, \\ b_h &= P_h * e^{t_h}. \end{aligned} \tag{1}$$

Among these, the selection of the anchor box adopts the method of dimensional clustering. Traditional clustering algorithms include hierarchical clustering and K-means clustering and model-based methods [36].

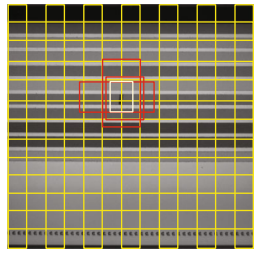

Figure 1: Schematic diagram of the prediction box of YOLOV3 in the 13 ∗ 13 cell.

YOLOV3 uses the K-means clustering algorithm to cluster the size of the target frame in the training set in order to obtain the optimal size of the anchor box. Thereby, a more accurate target frame can be predicted. The distance metric of the K-means clustering algorithm is given by

$$d(\text{box}, \text{centroid}) = 1 - \text{IOU}(\text{box}, \text{centroid}) \tag{2}$$

Here, box refers to the border size sample in the data set, and centroid refers to the cluster center size. The K-means clustering algorithm randomly selects K target points as the initial clustering center, and K represents K classifications. This random approach increases the randomness of the cluster and affects the clustering effect of the algorithm.

*2.2. SkNet.* In the neural network, the receptive fields of each layer have the same size, but in human vision, the size of the receptive fields will change depending on the size of the object. To make the neuron adaptively adjust the size of its receptive field for different sizes of input information, the selective kernel network (SkNet) [37] module is proposed. This module uses a nonlinear method to aggregate kernels of different sizes, and these kernels are mixed together via softmax attention. The size of the receptive field in different fusion layers is different, as shown in Figure 2.

SkNet is divided into three parts: split, fuse, and select. The first is the split operation. For the input $X$ ($C * h * w$, where $C$ is the dimension, $h$ is high, and $w$ is wide), the different receptive fields are obtained through the convolution kernels of 3 ∗ 3 and 5 ∗ 5, respectively. The two feature maps are $\tilde{U}$ and $\hat{U}$. Next is the fuse operation, which adds two feature maps to get $U$ as

$$U = \tilde{U} + \hat{U}. \tag{3}$$

To obtain global information, perform global average pooling operations:

$$s_c = F_{gp}(U_c) = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i, j). \tag{4}$$

At the same time, to improve the accuracy and adaptability of the network, a fully connected layer is added after the pooling layer:

$$z = F_{fc}(s) = \delta(\beta(W_s)), \tag{5}$$

where $\delta$ is Relu, $\beta$ is BN [38], and $z$ is equivalent to a queue [39] operation, that is, $z$ has a smaller dimension than $c$. The dimension of $z$ is set to $d$, and the value of $d$ is

$$d = \max(C/r, L). \tag{6}$$

Among them, $r$ and $L$ are artificially set, and $r$ is a ratio that compresses the dimensions. The select operation indicates that soft attention between channels is used to select information of different scales:

$$\begin{aligned} a_c &= \frac{e^{A_c z}}{e^{A_c z} + e^{B_c z}}, \\ b_c &= \frac{e^{B_c z}}{e^{A_c z} + e^{B_c z}}, \\ V_c &= a_c * \tilde{U}_c + b_c * \hat{U}, a_c + b_c = 1, \\ V &= [V_1, V_2, \cdots, V_C], V_c \in R^{H*W}. \end{aligned} \tag{7}$$

To summarize the idea of SkNet, there are several scale feature maps, and the features from squeeze are returned to c by several full connections, and then, the N fully connected results are put together. Then, perform softmax on each column vertically, so the same channel with different scales has different weights.

## 3. SK-FMYOLOV3

*3.1. FMYOLOV3*

*3.1.1. Predictive Box Clustering Algorithm with Fuzzy C-Means (FYOLOV3).* To reduce the randomness and improve the accuracy of the prediction frame, the clustering method is improved. The fuzzy clustering algorithm [40] is set to generate an initial clustering center. Then, the initial clustering center is passed into the K-means algorithm, and the result of the clustering is the initial position of the anchor.

First, the data are standardized. The set of the image classification objects is $X = x_1, x_2, \cdots, x_n$. In this set, there are $m$ indicators in any sample $x_j$, and the sample $x_j$ is used to label the characteristic index vector.

$$x_j = \left(x_{j,1}, x_{j,2}, \cdots, x_{j,m}\right). \tag{8}$$

In Eq. (3), $x_{j,m}$ represents the index of the $m$th characteristic in the sample $x_j$, and the matrix of the characteristic indicators of the $j * m$ samples is

$$\begin{bmatrix} x_{11}, & x_{12}, & \cdots, & x_{1m} \\ x_{21}, & x_{22}, & \cdots, & x_{2m} \\ \cdots, & \cdots, & \cdots, & \cdots \\ x_{j1}, & x_{j2}, & \cdots, & x_{jm} \end{bmatrix}, \tag{9}$$

constrain the value in $x_j$ to [0,1] through data transformation. The algorithm uses local neighborhood information and covariance to construct an objective function. The

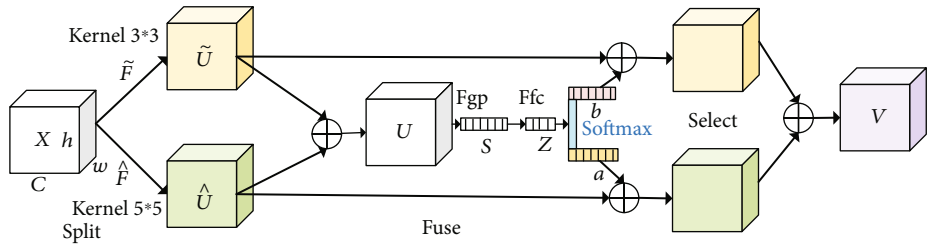


Figure 2: Adding a scale to the YOLOV3 model. The modified YOLOV3 can be used to fuse the feature maps of four scales.

objective function and constraints of the algorithm are as follows:

$$J_{FCM} = \sum_{j=1}^{N}\sum_{i=1}^{C} u_{ij}^{m} d^2\left(x_j, c_i\right) + \alpha_1 \sum_{n=1}^{N}\sum_{i=1}^{C} u_{ij}^{m} d^2\left(\bar{x}_j, c_j\right). \tag{10}$$

where $N$ is the total number of image pixels, $C$ is the number of image classifications, and $u_{ij}$ represents the degree of membership of the pixel $x_j$ belonging to the $i$ th category. $m$ is a fuzzy weighting coefficient greater than 1, and $c_i$ represents the $i$th cluster center. $d(x_j, c_i)$ represents the Mahalanobis distance from the $j$th data point to the $i$th cluster center, which is the covariance distance of the data. $\alpha_1$ represents the balance parameter that controls the influence of neighboring pixels. The Mahalanobis formula is as follows:

$$\begin{aligned} d^2\left(x_j, c_i\right) &= \left(x_j - c_i\right)^T V_i \left(x_j - c_i\right), \\ V_i &= |S|^{\frac{1}{p}} S^{-1}, \\ S &= \frac{\sum_{j=1}^{N}\sum_{i=1}^{C} u_{ij}^{m}\left(x_j - c_i\right)^T\left(x_j - c_i\right)}{\sum_{j=1}^{N}\sum_{i=1}^{C} u_{ij}^{m}}, \end{aligned} \tag{11}$$

where $|\bullet|$ represents the matrix determinant, and $p$ represents the dimension of the problem. Clustering center $c_i$ and membership $u_{ij}$ of the $i$th pixel can be obtained based on the Langer multiplier method:

$$\begin{aligned} c_i &= \frac{\sum_{j=1}^{N} u_{ij}^{m}\left(x_j + \alpha\bar{x}_j\right)}{(1+\alpha)\sum_{j=1}^{N} u_{ij}^{m}}, \\ u_{ij} &= \frac{\left[d^2\left(x_j, c_i\right) + \alpha d^2\left(\bar{x}_j, c_i\right)\right]^{1/m-1}}{\sum_{l=1}^{C}\left[d^2\left(x_j, c_l\right) + \alpha d^2\left(\bar{x}_j, c_l\right)\right]^{1/m-1}}, \end{aligned} \tag{12}$$

$c_i = c_1, c_2, c_3, \cdots, c_k$ is the initial cluster center of the K-means clustering algorithm, where $k$ is the number of categories. For each sample $x_i$ in the data set, calculate its distance to the $K$ cluster centers and divide it into the class corresponding to the smallest cluster center $c_i$. For each cluster center $c_i$, recalculate its cluster center:

$$c_i = \frac{1}{|c_i|}\sum_{x \in c_i} x \tag{13}$$

Repeat the distance calculation and update the distance center until the position of the cluster center no longer changes. The algorithm flow is presented in Algorithm 1.

*3.1.2. Multiscale Detection (MYOLOV3).* The YOLOV3 algorithm uses three different scale feature map fusions, using high resolution of low-level features and high semantic information of high-level features. By upsampling the features of different layers, objects are detected on three different scale feature layers. As shown in Figure 3, the bottom-level downsampling feature map is $13 * 13$, and the two upsampling feature maps are $26 * 26$ and $52 * 52$, respectively.

The YOLOV3 network has 32 times downsampling of the input detection image. The downsampling factor is high, the receptive field of the feature map is relatively large, and the shallow information is not fully utilized, resulting in some information loss after multilayer convolution. Therefore, this network is suitable for detecting large-sized objects in an image. In industrial production, the defects of objects are relatively small. For better detection of small defects, the original three-scale detection is extended to four-scale detection.

As shown in Figure 4, when multiscale fusion is performed, an upsampling fusion operation is used, a scale is added for the fusion operation, and a feature map with an upsampling of $104 * 104$ is added. Due to the addition of a scale, the anchor value also must be readjusted as shown in Table 1.

*3.2. SK-FMYOLOV3.* In industrial images with relatively small defects, the conventional YOLOV3 often has incorrect or missed defects. This is due to misidentification caused by the imbalance of the confidence distribution. To enable the network to learn the global features and autonomously improve the score of small defects, the SkNet structure is embedded in the improved YOLOV3 network. This makes the network make choices about information at different scales. Under the condition that the detection speed is guaranteed, the detection accuracy is improved, and the efficiency of real-time quality inspection in the industry is improved.

Considering that there is a $3 * 3$ convolution operation in the YOLOV3 convolution layer, there is also one in SkNet.

Input: image $x_j$, classification $c_i$
Step0: Standardize the input data, Eg. (3)(4), constrain the input value to [0,1].
Step1: Define the objective function and constraints of the algorithm, Eg. (5).
Step2: Obtain the cluster center $c_i$ and membership $u_{ij}$ of the i pixel by the Langer multiplier method.
Step3: Use $c_i = (c_1, c_2, c_3, ..., c_k)$ as the initial cluster center.
Step4: Calculate the distance from each sample $x_i$ to the cluster center according to the Eg. (6), (7).
Step5: Divide the sample $x_i$ into the class corresponding to the cluster center $c_i$ with the smallest distance.
Step6: Update the cluster center $c_i$ by the Eg. (8).
Step7: Repeat Step5 and Step6 until the value of $c_i$ is unchanged.
Output: output prediction box center $c_i$

Algorithm 1: The cluster box predicts flow that by fusing the fuzzy C-means algorithm into the K-means algorithm.

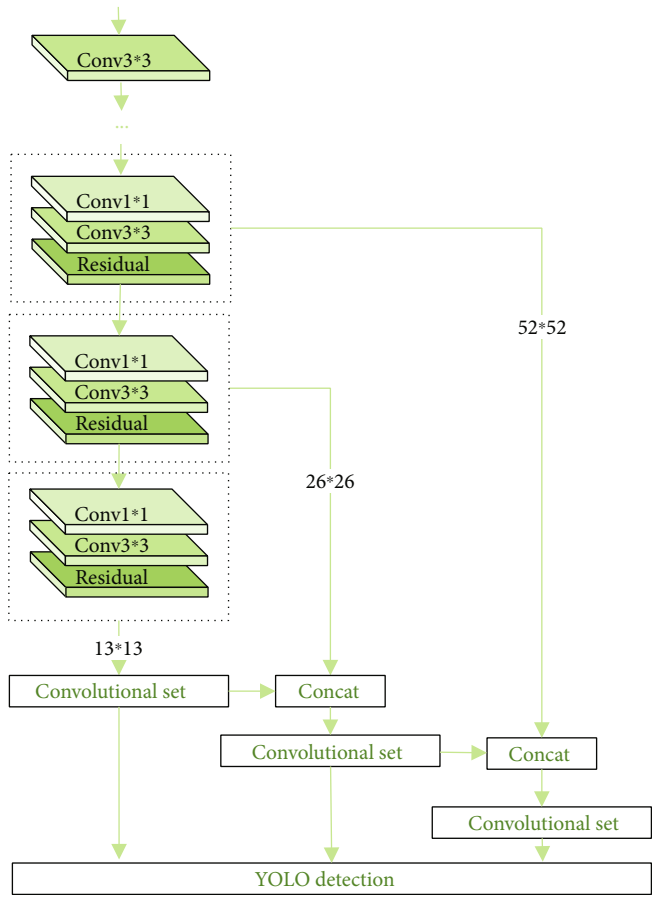


Figure 3: SkNet structure.

To maintain the original detection speed, starting from the original $3 * 3$ convolutional layer of layer 4 of YOLOV3, the subsequent $3 * 3$ convolutional layer was replaced with a SkNet structure. This makes the network have different receptive fields for feature maps of different sizes, replacing a total of 23 SkNet structures, as shown in Figure 5.

A feature map of $W * H * C$ is passed in a $1 * 1$ convolution layer, where $W$ is the width, $H$ is the height, and $C$ is the number of channels. Different receptive fields are obtained through the $3 * 3$ and $5 * 5$ convolution kernels, and the two feature maps are $\tilde{U}$ and $\hat{U}$, respectively. Add operations are performed on two feature maps to obtain $U$, and then, fuse and select operations are performed on $U$ to output feature maps. The specific parameter configuration of the designed SK-YOLOV3 is provided in Table 2.

SeNet (Squeeze and Excitation Networks) and SkNet are network structures proposed by the same team, and both of these introduce attention to improve the global receptive field. SeNet adaptively selects the channel, and SkNet adaptively selects the convolution kernel. Therefore, this paper also designed the SE-YOLOV3 network model for experimental comparison.

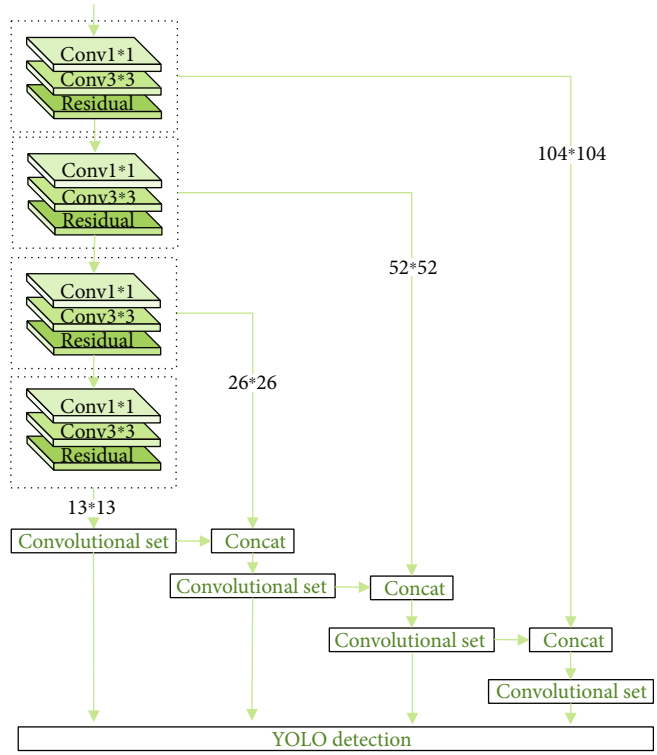


Figure 4: YOLOV3 network structure diagram. The original YOLOV3 contains three scales, and the prediction results of the yolo layer prediction box will be combined with the detection results of 3 scales.

Table 1: Initial value of the anchor box corresponding to improved multiscale feature map prediction.

| Feature layer | Feature map | Size of anchor |
|---|---|---|
| Layer 1 | 19 ∗ 19 | (103,197) (132,310) (301,346) |
| Layer 2 | 38 ∗ 38 | (61,195) (55,184) (78,247) |
| Layer 3 | 76 ∗ 76 | (50,74) (36,82) (30,61) |
| Layer 4 | 152 ∗ 152 | (11,19) (26,34) (18,25) |

## 4. Experiments

To accelerate the convergence speed of the network and avoid overfitting, 0.9 is used as the impulse constant, 0.0005 is used as the weight attenuation coefficient, and the initial learning rate is 0.0005. The experimental environment is the Ubuntu 14.04 operating system with Intel (R) Xeon (R) CPU E5-2698 v4 @ 2.20 GHz processor and 16 GB running memory (RAM), and the GPU is NVIDIA Tesla K80 with a 16 GB video memory.

The evaluation indicators are precision and recall. For these, precision is the precision rate that indicates the proportion of the samples in different categories that in fact belong to that category among the samples that are predicted to be positive:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{N}. \tag{14}$$

Here, 8 types of qualified test paper, "unqualified test paper," "Crooked," "Stains," "Marker pen," "Burr," "Short," and "Peeling" are used as detection targets, and predictions are made according to different categories. TP (True Positive) indicates the number of samples that correctly identify the defective target. FN (False Negative) indicates the number of samples for which no defective target was identified. FP (False Positive) indicates the number of samples that incorrectly identify a defective target. Recall is the recall rate that represents the ratio of the number of correctly detected targets to the total number of the targets in the test set:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{15}$$

The denominator of recall is true positives plus false negatives and represents the total number of samples.

## 5. Dataset

While the target objects in the images of large public datasets such as COCO [41] are relatively complete, there have been almost no studies of small defect classification for industrial products. Therefore, to verify the practicability of the algorithm, it is necessary to manually collect and create the data sets. (1) The urine test strip data are mainly obtained through high-definition camera shooting and crawler technology. The captured data are the main component, and the data obtained by the crawler technology are the minor component. A total of 1562 urine test strip images were collected

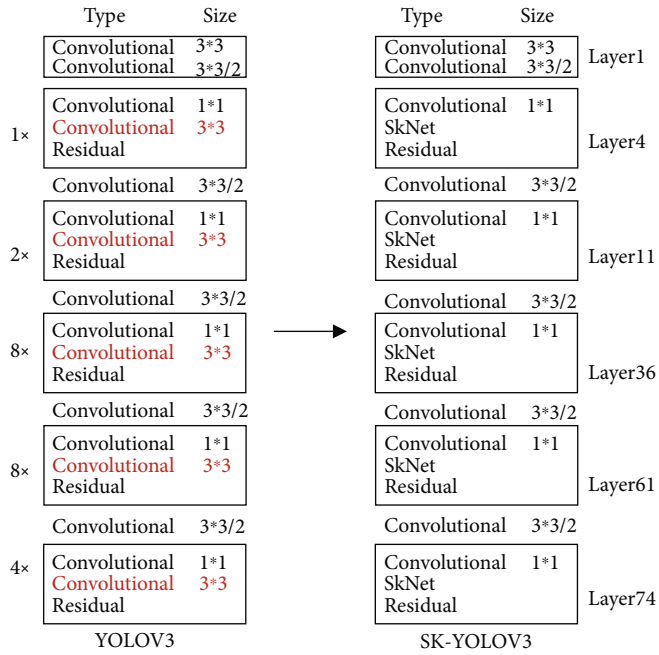


Figure 5: SK-FMYOLOV3 structure, replacing each $3 * 3$ convolutional layer with SkNet structure.

Table 2: SK-FMYOLOV3 parameter configuration. Here, $w$ is width, and $h$ is height. $x * y, z$ represents that the convolution kernel is $x * y$, and the number of filters is $z$. For example, for $3 * 3, 32$ represents a $3 * 3$ convolution kernel, and the number of filters is 32. $M$ is the number of branches, $G$ is the number of groups, and $r$ is the fully connected scaling scale.

| Output $(w * h)$ | YOLOV3 | SE–YOLOV3 | SK–FMYOLOV3 |
|---|---|---|---|
| $416 * 416$ 1 | $3 * 3, 32$ | $3 * 3, 32$ | $3 * 3, 32$ |
| $208 * 208$ | $3 * 3/2, 64$ | $3 * 3/2, 64$ | $3 * 3/2, 64$ |
| $208 * 208$ | $\begin{bmatrix} 1 * 1, 32 \\ 3 * 3, 64 \end{bmatrix} \times 1$ | $\begin{bmatrix} 1 * 1, 32 \\ 3 * 3, 64 \\ fc, (16, 64) \end{bmatrix} \times 1$ | $\begin{bmatrix} 1 * 1, 32 \\ \mathrm{SK}(M = 2, G = 32, r = 16) \end{bmatrix} \times 1$ |
| $104 * 104$ | $\begin{bmatrix} 1 * 1, 64 \\ 3 * 3,128 \end{bmatrix} \times 2$ | $\begin{bmatrix} 1 * 1, 64 \\ 3 * 3,128 \\ fc, (32,128) \end{bmatrix} \times 2$ | $\begin{bmatrix} 1 * 1, 64 \\ \mathrm{SK}(M = 2, G = 32, r = 16) \end{bmatrix} \times 2$ |
| $52 * 52$ | $\begin{bmatrix} 1 * 1,128 \\ 3 * 3,256 \end{bmatrix} \times 8$ | $\begin{bmatrix} 1 * 1,128 \\ 3 * 3,256 \\ fc, (64,256) \end{bmatrix} \times 8$ | $\begin{bmatrix} 1 * 1,128 \\ \mathrm{SK}(M = 2, G = 32, r = 16) \end{bmatrix} \times 8$ |
| $26 * 26$ | $\begin{bmatrix} 1 * 1,256 \\ 3 * 3,512 \end{bmatrix} \times 8$ | $\begin{bmatrix} 1 * 1,256 \\ 3 * 3,512 \\ fc, (128,512) \end{bmatrix} \times 8$ | $\begin{bmatrix} 1 * 1,256 \\ \mathrm{SK}(M = 2, G = 32, r = 16) \end{bmatrix} \times 8$ |
| $13 * 13$ | $\begin{bmatrix} 1 * 1,512 \\ 3 * 3, 1024 \end{bmatrix} \times 4$ | $\begin{bmatrix} 1 * 1,512 \\ 3 * 3, 1024 \\ fc, (256, 1024) \end{bmatrix} \times 4$ | $\begin{bmatrix} 1 * 1,512 \\ \mathrm{SK}(M = 2, G = 32, r = 16) \end{bmatrix} \times 4$ |

with a resolution of $2456 * 2058$ pixels. (2) The following methods are used for data enhancement of the original image: magnification (image width and height are enlarged to 1.5 times), reduction (image width is reduced to 0.3, height is reduced to 0.5, and image size is guaranteed to be a multiple of 32), brightness enhancement and reduction, flipping (90° and 180°), and clipping. Finally, 11634 images were obtained. (3) Labeling was used to mark 8 kinds of urine test

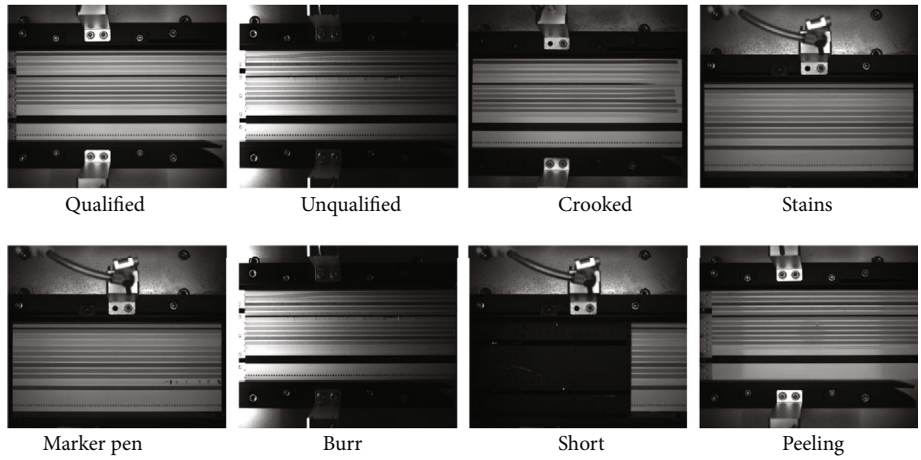


Figure 6: A classification example of the urine test strip database constructed in this work.

paper defects in 1562 images. As shown in Figure 6, the images are classified as "qualified test paper," "unqualified test paper," "Crooked," "Stains," "Marker pen," "Burr," "Short," and "Peeling." The specific method is the selection of all of the defects in the image box and obtain the XML file in the VOC format. (4) The xml file is converted to a txt file with the format of 'tag' + 'X' + 'Y' + 'W' + 'H', and 6,634 images are randomly selected as the training set and 5000 are selected as the test set.

*5.1. SK-FMYOLOV3 Convergence Verification.* Based on the improved YOLOV3 structure, the SkNet structure is embedded to train on a homemade urine test strip dataset. Iterative training on the GPU server 300 times, the results show that the model can quickly converge to a stable state during the training process. During the training process, log information is collected for each iteration of the SK-FMYOLOV3 model training. GIoU [42] is used as the loss of the detection task, and the objectness [43] is recorded during the training process, as well as the val GIoU and val objectness of the verification set. GIoU takes into account the nonoverlapping areas that IOU does not take into account and can reflect the manner in which the predicted box and the ground truth overlap. The objectness value represents the probability of the target in the prediction box. Through the visualization of information, during the training process, as the number of iterations continues to increase, the loss function gradually converges in the first 200 iterations. The GIoU value of the training set and the validation set is stable at approximately 1.15, and the objectness value is stable at approximately 1.0, as shown in Figure 7.

*5.2. Impact of Different Improvement Strategies on the Prediction Box.* The impact on the accuracy of the prediction box is calculated for the three improved strategies proposed above, using the original YOLOV3 as the reference, as shown in Table 3. As shown in Table 3, FYOLOV3 represents the addition of an improved prediction box clustering algorithm based on the YOLOV3 algorithm. MYOLOV3 stands for multiscale improved YOLOV3 algorithm. FMYOLOV3 represents the addition of an improved predictive box clustering algorithm and an improved multiscale algorithm based on the YOLOV3 algorithm. SK-FMYOLOV3 represents the addition of SkNet structure on the basis of FMYOLOV3.

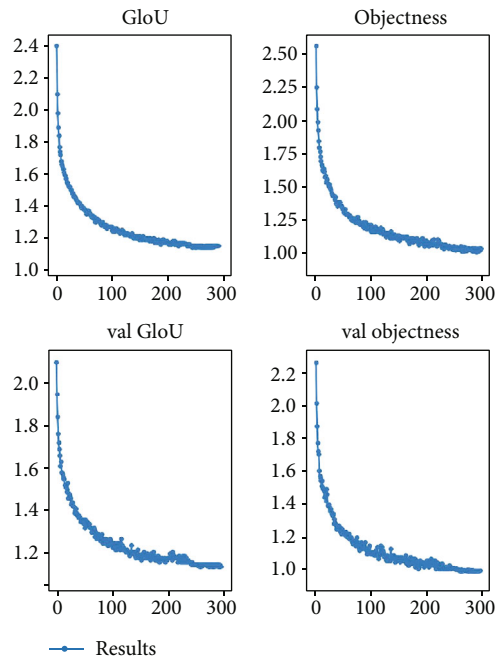


Figure 7: Loss convergence diagram of model training.

Table 3: Impact of different improvement strategies on model accuracy.

| Strategy | Improved prediction box | Multiscale | SkNet | Average IOU |
|---|---|---|---|---|
| YOLOV3 | ✗ | ✗ | ✗ | 78.56 |
| FYOLOV3 | ✓ | ✗ | ✗ | 84.34 |
| MYOLOV3 | ✗ | ✓ | ✗ | 82.23 |
| FMYOLOV3 | ✓ | ✓ | ✗ | 85.09 |
| SK-FMYOLOV3 | ✓ | ✓ | ✓ | **86.57** |

Table 4: Performance comparison of different detection models.

| Model | Improved prediction box | Multiscale | SkNet | $P$ (%) | $R$ (%) | Speed/ms |
|---|---|---|---|---|---|---|
| R-CNN [21] | ✗ | ✗ | ✗ | 78.5 | 64.2 | 1200 |
| FAST-RCNN [23] | ✗ | ✗ | ✗ | 86.6 | 79.7 | 700 |
| FASTER-RCNN [24] | ✗ | ✗ | ✗ | 89.8 | 87.3 | 350 |
| YOLOV3 [32] | ✗ | ✗ | ✗ | 87.6 | 71.4 | **230** |
| SE-YOLOV3 | ✗ | ✗ | ✗ | 92.5 | 86.1 | 350 |
| FYOLOV3 | ✓ | ✗ | ✗ | 93.1 | 90.2 | 310 |
| MYOLOV3 | ✗ | ✓ | ✗ | 92.3 | 89.6 | 340 |
| FMYOLOV3 | ✓ | ✓ | ✗ | 94.6 | 91.7 | 360 |
| SK-FMYOLOV3 | ✓ | ✓ | ✓ | **96.8** | **94.5** | 390 |

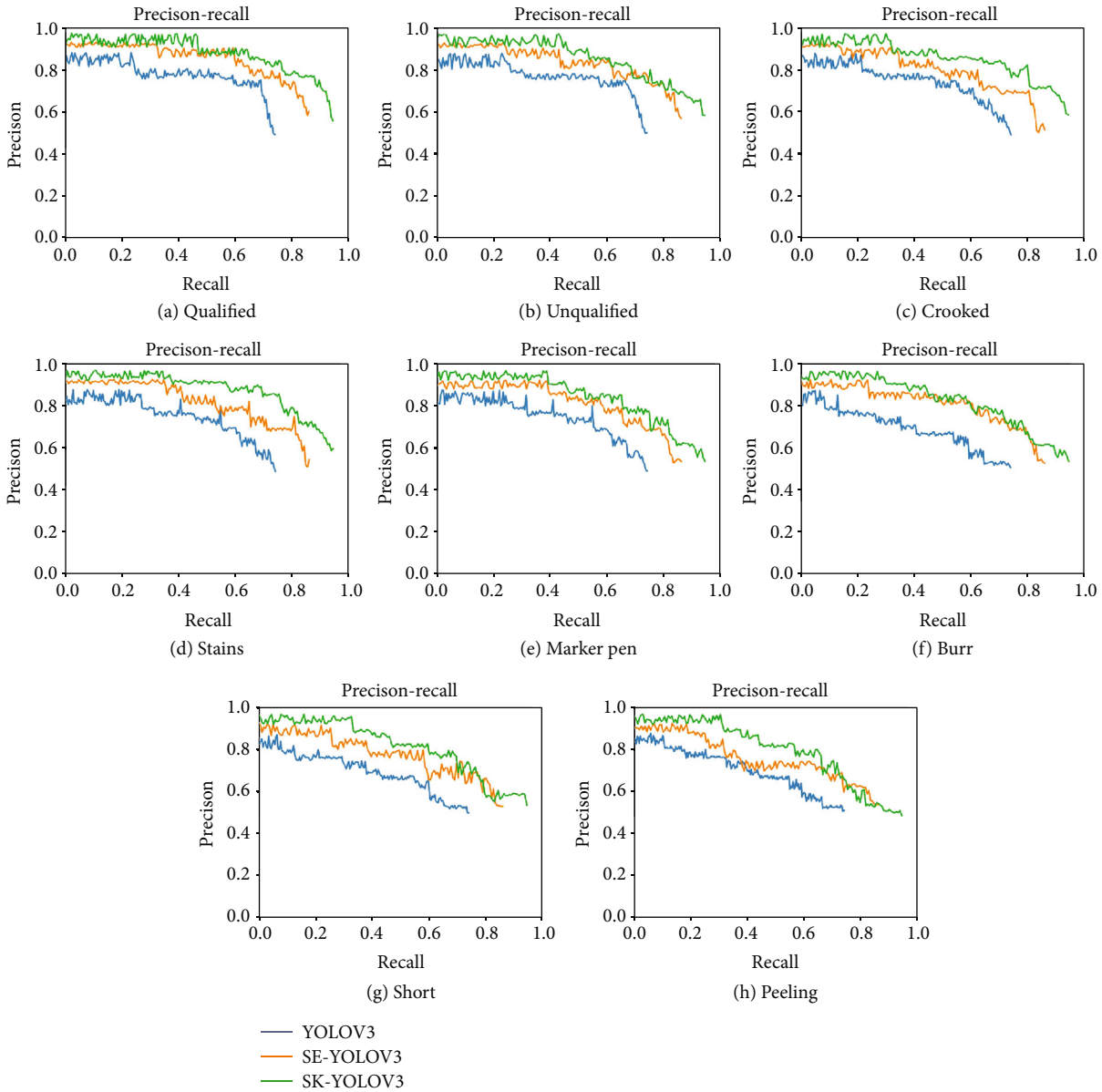


Figure 8: Urine test paper defect detection rate and recall rate are divided into 8 categories: qualified test paper, unqualified test paper, Crooked, Stains, Marker pen, Burr, Short, and Peeling.

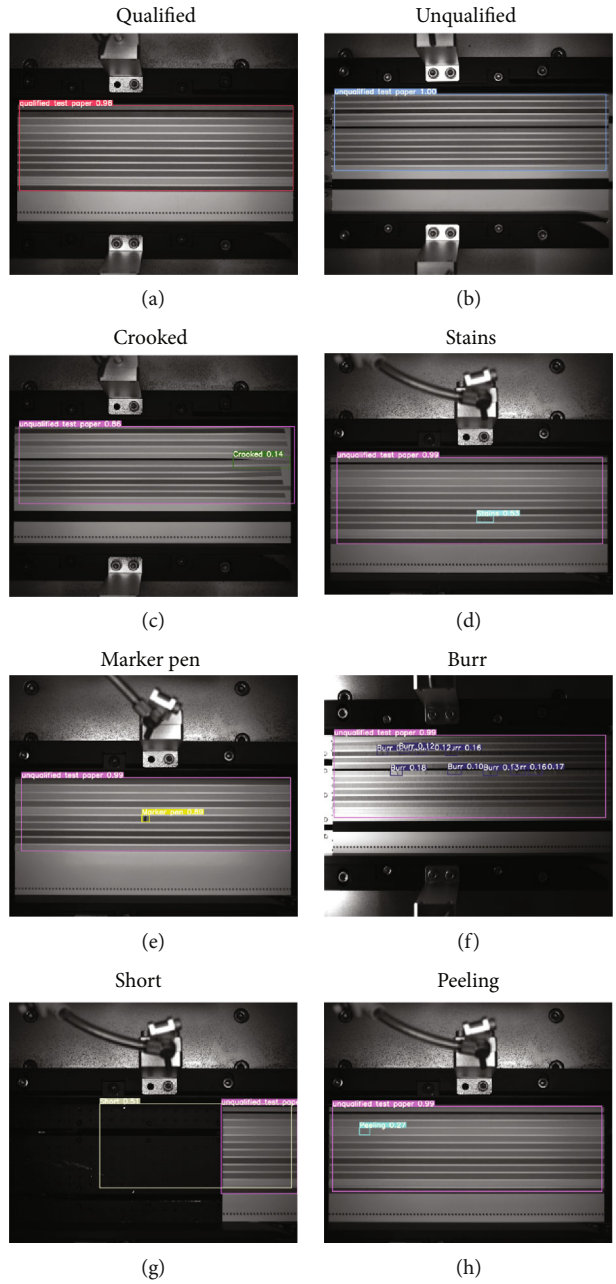


Figure 9: SK-YOLOV3 results of urine defect detection. The accuracy rate for the classification of qualified products and unqualified products is above 95, showing that the defect categories in the unqualified products can be accurately identified.


Figure 10: Test results of the qualified products for SK-YOLOV3, YOLOV3, and Faster-RCNN.

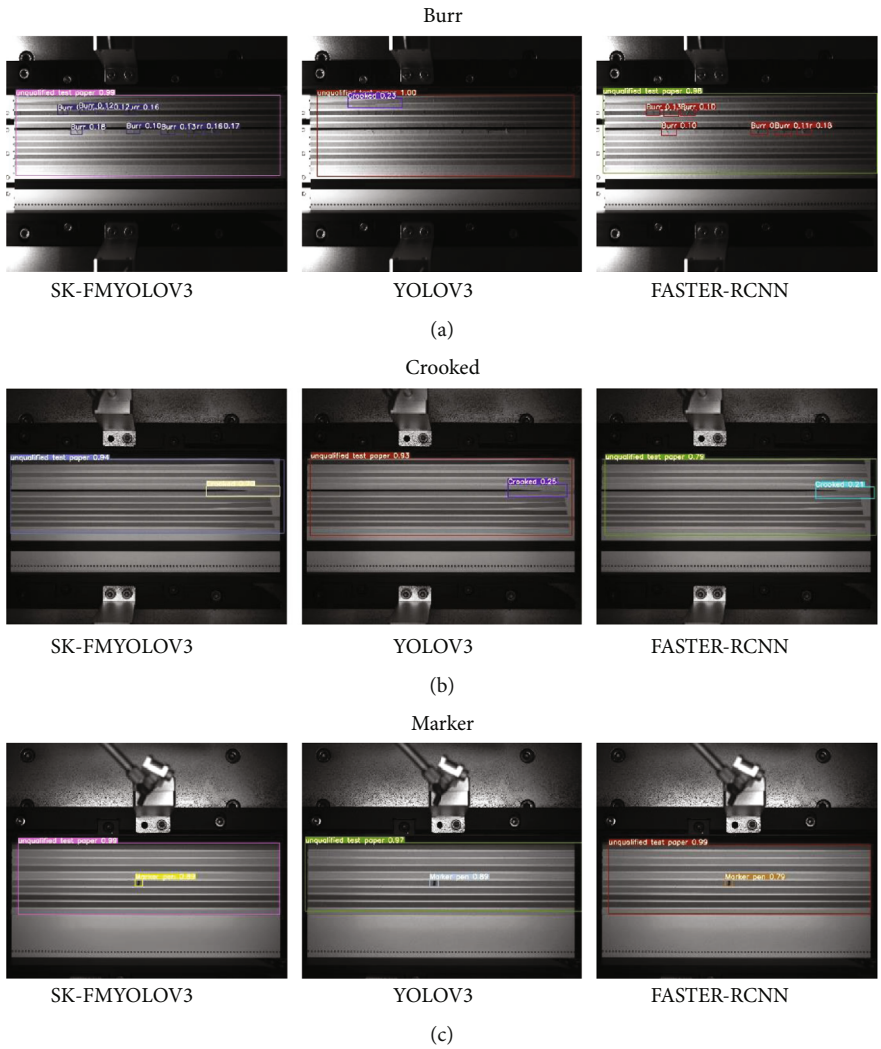


Figure 11: Test results of the unqualified products (burr, crooked, and marker) for SK-YOLOV3, YOLOV3, and Faster-RCNN.

Each improvement strategy used in this paper improves the performance of the original YOLOV3 detection network to varying degrees. Among these, the improvement of the prediction box clustering algorithm displays the most significant improvement in the model accuracy, and the average IOU has increased by nearly 6 percentage points. The improvement of the multiscale algorithm leads to the average IOU increase of nearly 4 percentage points. The improved clustering box prediction algorithm and multiscale algorithm based on YOLOV3 increased the average IOU by nearly 7 percentage points. Combining all of the improvement strategies, the final average IOU is improved by nearly 8 percentage points over the original YOLOV3 network.

*5.3. Performance Evaluation.* The accuracy rate ($P$) and recall rate ($R$) are used as evaluation indicators, and the same data set is used in the same experimental environment. The YOLOV3 method with different improvement strategies is compared with R-CNN, fast RCNN, and faster RCNN. The test results of the qualified and unqualified products are shown in Table 4.

As seen from the above table, the accuracy and recall of the SK-FMYOLOV3 network is the highest. The reason is that the SkNet structure allows feature maps to be selected for training by different convolution kernels, improving the score of small features. At the same time, the accuracy of classification is increased, making the accuracy and recall of the network higher. The fastest network is YOLOV3, because the improved algorithm increases the number of the layers in the network and therefore increases the recognition time. FYOLOV3 is better than MYOLOV3, because the prediction box clustering algorithm is added to avoid the impact of random initial points on the prediction result. MYOLOVS is better than YOLOV3, because the annotations in the data set are small defects. After adding a small-scale feature fusion, the recall rate is improved, so that previously unrecognized defects are identified. SK-FMYOLOV3 achieved the highest recall and precision, because the convolution autonomous selection can be trained according to the size of different feature maps, leading to an improvement in the accuracy of classification. By embedding SkNet in the improved YOLOV3 structure, the accuracy rate is increased by 9 percentage points and the recall rate is increased by 23 percentage points.

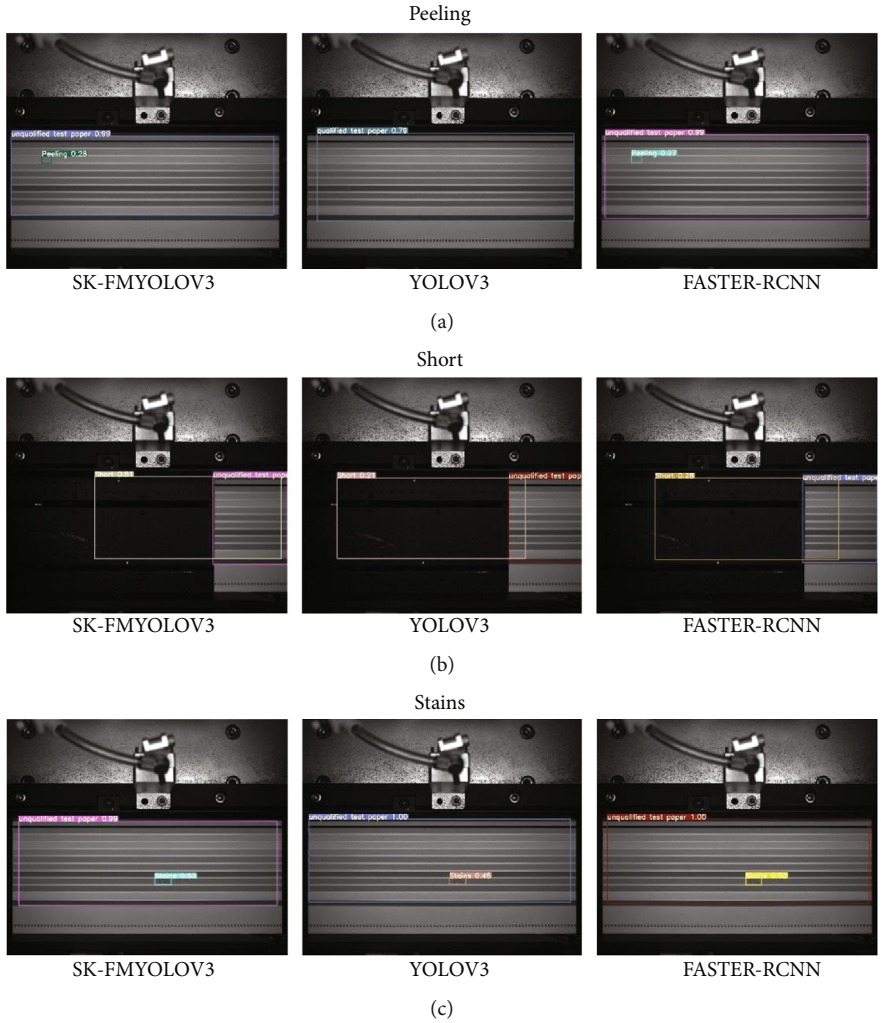


Figure 12: Test results of the unqualified products (peeling, short, and stains) for SK-YOLOV3, YOLOV3, and Faster-RCNN.

The accuracy and recall of the 8 classifications of the homemade urine test strip data set in SK-FMYOLOV3 are shown in Figure 8. In the first 300 iterations, a higher precision is observed, but as the number of iterations increases, overfitting will occur, the recall will increase, and the precision will decrease. The test results of SK-FMYOLOV3 for 8 categories are as shown in Figure 9.

*5.4. Comparison of Experimental Results.* In the same experimental environment, the same number of iterations are used for training (epoch = 300). The test results of this method and faster-RCNN, YOLOV3 on the homemade urine test strip data set are shown below.

The first is the detection of qualified products, as shown in Figure 10. The accuracy rate of the qualified urine test paper in this method is 0.99, YOLOV3 is 0.73, and that of faster-RCNN is 0.89; thus, this method is more effective for the detection of qualified products.

Then, it is the defect detection of nonconforming products, as shown in Figure 11. For the detection of burr, it is observed that SK-FMYOLOV3 detected 11 burrs on the urine test strip. YOLOV3's attention to small defects is not as high as that to large defects. Therefore, crooked was detected on urine test strips, and no burr was detected. Faster-RCNN detected 7 burrs on urine test strips. For the detection of crooked defects as shown in Figure 11, all three algorithms were successful. Because Crooked defects are relatively large and the features are obvious, all three algorithms show better performance. As shown in Figure 11, for detection of marker defects, the three algorithms can detect the marker better, and the marker is also a relatively obvious defect.

The detection of peeling defects is shown in Figure 12. It is observed that SK-FMYOLOV3 and Faster RCNN can detect these defects, while YOLOV3 cannot detect these defects. As shown in Figure 12, for the detection of short defects, it is found that the three algorithms show good performance, but the accuracy of the algorithm in this paper is approximately 0.5, while those for the other two models are approximately 0.2. The detection of stains defects is shown in Figure 12. All three algorithms can be detected to carry out the detection. The accuracy of this algorithm and of

Faster-RCNN is approximately 0.5, and that of YOLOV3 is 0.45. It is important to note that in addition to the detection of defects, each urine test strip also detects the nonconforming products. As can be observed from the above group of figures, the accuracy rate of unqualified products is approximately 0.99, and showing that the method is suitable for use in the classification of qualified and unqualified products of industrial products.

## 6. Conclusions

To solve the problem of detection accuracy of defective products in industrial production, this paper proposes an SK-FMYOLOV3 algorithm based on the YOLOV3 network. First, fuzzy mean clustering is used to generate the initial clustering points to avoid the influence of randomly initialized prediction frame on detection accuracy. Then, the original three-scale prediction is changed to four scales, making the algorithm more suitable for detecting smaller defects. Finally, the SkNet structure is merged, so that the feature map selects the appropriate convolution kernel for training through the attention mechanism, and the scores of the defects that are not easy to identify are higher. The proposed network structure is based on the homemade urine test strip data set, and the detection precision rate and recall rate of the qualified urine test strip and the unqualified urine test strip are 96.8 and 94.5, respectively, were obtained. This method can accurately identify the 6 types of small defects in nonconforming products. In future research, we will consider using the network structure for other industrial products to conduct experiments in order to save human resources and improve production efficiency.

## Data Availability

The [Urine dipstick dataset] data used to support the findings of this study were supplied by [Rui Yang] under license and so cannot be made freely available. Requests for access to these data should be made to [Rui Yang, 792481404@qq.com].

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 61772149, U1701267, and 61762028), GUET Excellent Graduate Thesis Program (No. 18YJPYSS15), Guangxi Key Laboratory of Image and Graphic Intelligent Processing Project (No. GIIP2003), and Guangxi Science and Technology Project (Nos. AB20238013, ZY20198016, 2018GXNSFAA294127).

## References

[1] Y. Min, B. Xiao, J. Dang, B. Yue, and T. Cheng, "Real time detection system for rail surface defects based on machine vision," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, Article ID 3, 2018.

[2] R. Lan, Y. Zhou, Z. Liu, and X. Luo, "Prior knowledge-based probabilistic collaborative representation for visual recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 4, pp. 1498–1508, 2020.

[3] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "Madnet: a fast and lightweight network for single-image super resolution," *IEEE Transactions on Cybernetics*, vol. 99, pp. 1–11, 2020.

[4] R. C. Hardie, R. Ali, M. S. De Silva, and T. M. Kebede, "Skin lesion segmentation and classi_cation for isic 2018 using traditional classifiers with hand-crafted features," https://arxiv.org/abs/1807.07001.

[5] C. Yu, G. Zhang, and Y. Gao, "Improved threshold-based segmentation method for millimeter wave radiometric image," in *Proceedings of the 2019 International Conference on Modeling, Simulation, Optimization and Numerical Techniques (SMONT 2019)*, Shenzhen guangdong, China, 2019.

[6] H. Lyu, H. Fu, X. Hu, and L. Liu, "Esnet: edge-based segmentation network for real-time semantic segmentation in traffic scenes," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1855–1859, Taipei, Taiwan, 2019.

[7] R. Priyadharsini and T. S. Sharmila, "Object detection in underwater acoustic images using edge based segmentation method," *Procedia Computer Science*, vol. 165, pp. 759–765, 2019.

[8] Z. Zhou, Q. M. J. Wu, Y. Yang, and X. Sun, "Region-level visual consistency verification for large-scale partial-duplicate image search," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 2, pp. 1–25, 2020.

[9] Z. Zhou, Y. Mu, and Q. M. J. Wu, "Coverless image steganography using partial-duplicate image retrieval," *Soft Computing*, vol. 23, no. 13, pp. 4927–4938, 2019.

[10] M. Elawady, C. Ducottet, O. Alata, C. Barat, and P. Colantoni, "Wavelet-Based Reflection Symmetry Detection via Textural and Color Histograms: Algorithm and Results," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1734–1738, Venice, Italy, 2017.

[11] Z. Xing and H. Jia, "Multilevel color image segmentation based on glcm and improved salp swarm algorithm," *IEEE Access*, vol. 7, pp. 37672–37690, 2019.

[12] H. Dong, X. Zhang, Y. Guo, and F. Wang, "Deep multi-scale gabor wavelet network for image restoration," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2028–2032, Barcelona, Spain, 2020.

[13] D. N. Thanh, N. N. Hien, V. S. Prasath, U. Erkan, and A. Khamparia, "Adaptive thresholding skin lesion segmentation with gabor filters and principal component analysis," in *Intelligent Computing in Engineering*, Advances in Intelligent Systems and Computing, V. Solanki, M. Hoang, Z. Lu, and P. Pattnaik, Eds., pp. 811–820, Springer, Singapore, 2020.

[14] M. M. T. Zadeh, M. Imani, and B. Majidi, "Fast facial emotion recognition using convolutional neural networks and gabor filters," in *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, pp. 577–581, Tehran, Iran, 2019.

[15] J. Nazarinezhad and M. Dehghani, "A contextual-based segmentation of compact polsar images using markov random field (mrf) model," *International Journal of Remote Sensing*, vol. 40, no. 3, pp. 985–1010, 2019.

[16] W. Li, Y. Chen, W. Sun et al., "A gingivitis identification method based on contrast-limited adaptive histogram equalization, gray-level co-occurrence matrix, and extreme learning

machine," *International Journal of Imaging Systems and Technology*, vol. 29, no. 1, pp. 77–82, 2019.

[17] Y. Huang, S. Xu, L. Yang, S. Zhao, Y. liu, and Y. Shi, "Defect detection during laser welding using electrical signals and high-speed photography," *Journal of Materials Processing Technology*, vol. 271, pp. 394–403, 2019.

[18] B. Li, F. Zhao, Z. Su, X. Liang, Y. K. Lai, and P. L. Rosin, "Example-based image colorization using locality consistent sparse representation," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5188–5202, 2017.

[19] Y. Wang, Z. Cai, Z. H. Zhan, Y. J. Gong, and X. Tong, "An optimization and auction-based incentive mechanism to maximize social welfare for mobile crowdsourcing," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, pp. 414–429, 2019.

[20] R. Lan, H. Lu, Y. Zhou, Z. Liu, and X. Luo, "An LBP encoding scheme jointly using quaternionic representation and angular information," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4317–4323, 2020.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, 2014.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[23] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Santiago, Chile, 2015.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[25] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, pp. 379–387, Morgan Kaufmann, 2016.

[26] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *Computer Vision-ECCV 2016. ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9905 of Lecture Notes in Computer Science, pp. 21–37, Springer, Cham, 2016.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Las Vegas, NV, USA, 2016.

[28] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7263–7271, Honolulu, HI, USA, 2017.

[29] N. S. Samarawickrama, *Faster R-CNN Based CubeSat Close Proximity Detection and Attitude Estimation*, Mississippi State University, 2019.

[30] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: deconvolutional single shot detector," https://arxiv.org/abs/1701.06659.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.

[32] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," https://arxiv.org/abs/1804.02767.

[33] J. Tao, H. Wang, X. Zhang, X. Li, and H. Yang, "An object detection system based on yolo in traffic scene," *2017 6th International Conference on Computer Science and Network Technology (ICCSNT)*, 2017, pp. 315–319, Dalian, China, 2017.

[34] J. Deng, G. Pang, Z. Zhang, Z. Pang, H. Yang, and G. Yang, "cgan based facial expression recognition for human-robot interaction," *IEEE Access*, vol. 7, pp. 9848–9859, 2019.

[35] D. Weicong, J. Longxu, L. Guoning, and Z. Zhiqiang, "Real-time airplane detection algorithm in remote-sensing images based on improved yolov3," *Opto-Electronic Engineering*, vol. 45, no. 12, article 180350, 2018.

[36] H. A. Taboada and D. W. Coit, "Data clustering of solutions for multiple objective system reliability optimization problems," *Quality Technology & Quantitative Management*, vol. 4, no. 2, pp. 191–210, 2016.

[37] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 510–519, Long Beach, California, USA, 2019.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," https://arxiv.org/abs/1502.03167.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, 2018.

[40] K. Nongmeikapam, W. K. Kumar, and A. D. Singh, "Fast and automatically adjustable grbf kernel based fuzzy c-means for cluster-wise coloured feature extraction and segmentation of mr images," *IET Image Processing*, vol. 12, no. 4, pp. 513–524, 2018.

[41] Y. Dong, H. Su, B. Wu et al., "Efficient decision-based black-box adversarial attacks on face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7714–7722, Long Beach, CA, USA, 2019.

[42] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: a metric and a loss for bounding box regression," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2020.

[43] M. M. Cheng, Y. Liu, W. Y. Lin, Z. Zhang, P. L. Rosin, and P. H. S. Torr, "BING: Binarized Normed Gradients for Objectness Estimation at 300fps," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014.