Journal of Computers Vol. 34 No. 4, August 2023, pp. 53-68
doi: 10.53106/199115992023083404005

X-ray Image Prohibited Item Detection Algorithm Based on Improved PP-YOLO

Ji-Kai Zhang[1], Yue Liu[1], Xiao-Qi Lv[2], Yong Liang[1*]

[1] Department of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China
jkzhang0314@imust.edu.cn, 13190710227@163.com, 13088906116@163.com

[2] Department of Information Engineering, Inner Mongolia University of Technology, Hohhot 010051, China
lxiaoqi@imust.edu.cn

Received 19 July 2022; Revised 7 November 2022; Accepted 4 January 2023

Abstract. In order to solve the problems of missing detection due to overlap and occlusion of contraband in X-ray images and low accuracy of small object detection, we propose a single-stage object detection framework based on PP-YOLO. Compared with the traditional prohibited item detection algorithm, it adds CBAM module on the basis of ResNet50 feature extraction network to enhance the feature extraction ability; For increasing the detail features of the detection layer, MSF module is introduced into FPN, which fuses the feature map with accurate position information in the lower layer and the feature map with strong semantic information in the higher layer; The partial convolution of backbone is improved to CompConv to accelerate the processing speed of the model, which compresses the network structure and improves the inference speed without losing performance. The results show that the mAP of the improved network for prohibited item detection is 94.67%, and the processing speed reaches 45 FPS, which means that the recognition accuracy and reasoning speed of this method have been improved to some extent.

Keywords: object detection, prohibited item detection, X-ray image, PP-YOLO, attention echanism

1 Introduction

Although X-ray contraband detection has been widely used in the security detection process of public transportation and key public places, the manual detection is still the main process of X-ray image detection, which not only has low detection efficiency, but also inevitably leads to problems such as wrong detection and missing detection, in order to reduce the problems of false and missing detection of contraband in the security inspection process, this paper studies these problems in the security inspection process, so as to make travel safer and more reliable. Recently, under the development of computer software and hardware and deep learning technology, detection technology based on computer vision has been widely used in industry, agriculture, security and other fields and achieved good results. Some scholars have also applied it to prohibited item detection, among which Turcsany et al. [1] applied the classical BoVW model, which had a series of feature point detectors and descriptors, and was supported by support vector machine and random forest, and achieved more accurate detection results on large X-ray baggage image datasets; According to the shape information of contraband, Literature [2] proposed to use implicit shape model for threat detection, which was based on visual vocabulary and a generation structure. The detection effect was better on the dataset including blades, darts and pistols, but the dataset used was simple and cannot verify the recognition effect in real scenes; Multi-view detection [3, 4] can provide more abundant information than single view detection. Multi-view detection can suppress false alarm information and provide more accurate detection results, but the design of multi-view system is more complex; Aiming at the subjectivity, missed judgment and misjudgment in manual detection, Wang [5] and others effectively distinguished different kinds of contraband based on Tamura texture and random forest X-ray picture foreign body classification method in 2017; X-ray images are different from ordinary images, there are some problems such as unclear edges and cluttered background. In order to solve the problem of unclear contour of the detected object, Literature [6] proposed a foreground and background segmentation method based on color information. The recognition accuracy rate is 77% in a dataset containing 32253 subway security detection images by a deep convolutional neural networks (DCNNs) based on Faster-RCNN.

* Corresponding Author

Factors such as occlusion and overlap of objects in X-ray images are a major obstacle to object detection. In order to solve such problems, Galvez et al. [7] proposed a object detector based on YOLO [8] in 2019 for prohibited item detection, and found that training detectors from scratch is better than transfer learning; Small object detection has always been a difficult problem in object detection. With the purpose of improve the detection accuracy of small objects, Ji et al. [9] proposed a feature fusion object detection algorithm for image sub-regional detection in 2019. On the basis of SSD (Single Shot Multibox Detector) detector, shallow and deeper features were fused to improve the receptive field of shallow feature map and improve the detection effect of small instances; For the same purpose, Literature [10] established an automatic detection model based on feature fusion FSSD system, and used hole convolution semantic enrichment module (SEM) to extract low-level features and fuse high-level features to achieve effective detection; In 2021, Guo [11] and others proposed YOLO-C detection network on the basis of YOLOv3, which optimized the feature extraction ability of the model through feature enhancement module, and enhanced the linear expression ability of the model by feature fusion, and finally achieved good detection for small objects.

Some scholars have applied image segmentation technology to the field of object detection, and achieved good detection results. Chouai et al. [12] combined DCNNs with antagonistic self-encoder as a powerful feature extractor, and segmented the image into possibly overlapping areas and organic and inorganic images. According to different area information, better prohibited item detection results have been achieved; Literature [13] applied semantic segmentation to detect contraband multi-targets, and empty pyramid convolution module to improve the ability of multi-scale information mining, but the detection time was too long to meet the real-time requirements.

Different materials have different colors in X-ray imaging. Based on the idea of material classification, Literature [14] proposed XMC RCNN network to separate overlapping X-ray images and obtained renderings of different materials, which achieved higher detection accuracy; Based on the contour information of contraband, Hassan et al. [15] introduced to use the object boundary-driven framework for processing in 2020, and transmitted the generated contour suggestions to the neural network for recognition, which achieved good recognition effect, but the boundary contour was unclear in the case of overlap and the recognition effect was poor.

To sum up, there are many factors in the segmentation of X-ray images, such as high target overlap and uneven scales, which result in poor detection accuracy of contraband targets, dithering detection accuracy and high miss rate of small targets. Three improvements have been made to solve these problems:

(1) Add attention mechanism to improve the performance of feature extraction network.

(2) The MSF module is introduced into the FPN to fuse the high and low level features, and the pr-ecise position information is integrated into the higher level features.

(3) The partial convolution of the backbone is improved to CompConv (CC) to speed up the model processing.

The organizational structure of the article is as follows:

The first part is the introduction, which briefly introduces the research methods in the detection of contraband and some problems that need to be solved.

The second part is related work, which reviews the research work done by previous researchers and introduces the relevant technologies used in this paper.

The third part introduces the network structure of this paper, including feature extraction network, feature fusion module, detection head and other components.

The fourth part analyzes and improves the network model, proposes the improvement methods according to the existing problems, and introduces the functions of the improved parts one by one.

The fifth part is the data set and data processing part used in the experiment. In the data processing part, a new data generation method is proposed.

The sixth part is the analysis of the experimental results. Comparative experiments are carried out on several networks, and the effectiveness of the proposed improvements is verified.

Finally, the conclusion part summarizes the proposed network, summarizes the limitations of the current work, and looks forward to the future research direction.

2 Related Work

2.1 X-ray Prohibited Item Detection

Object detection is one of the three major tasks of computer vision, which is usually described as a classification problem based on sliding window in the early days, object detection. With the rapid development of computer hardware and the rise of neural networks, CNN-based methods gradually replace the early detection methods. This method can be subdivided into two categories: one is based on regional proposal, such as RCNN [16-18] series and other typical double-stage detection networks. The other type does not need regional proposals, such as YOLO [19-22] series and SSD [23], Retina-Net [24] and other single-stage detection networks. The method based on RPN usually has higher accuracy than the method without proposal, but the latter has higher detection speed, so the two methods have their own advantages. Literature [25-27] applied a typical single-stage detection network to detect dangerous articles, which was improved on the premise of ensuring the detection speed and improving the detection accuracy of contraband. Wu et al. [28] improved the double-stage detection network Faster-RCNN to realize accurate detection of contraband. Some scholars have applied image segmentation technology to the detection of prohibited item. Literature [29] introduced semantic segmentation to extract and classify the region of interest in X-ray, and finally realized the effective recognition of contraband. However, such methods consume more computing resources and are difficult to deploy in practice.

2.2 Feature Fusion

In the process of feature extraction, feature images with different resolutions contain different semantic information. The feature map with higher resolution has more accurate position information, while the feature map with lower resolution contains abundant semantic information although the detail information is lost seriously. Therefore, the fusion of features with different resolutions can effectively improve the detection accuracy. Literature [30-34] used feature fusion to enhance the semantic information contained in the feature map and improved the accuracy of detection and segmentation tasks.

Inspired by the above ideals, we proposes a single-stage detection network based on PP-YOLO network, which is suitable for X-ray contraband recognition. Attention mechanism and feature fusion technology are used to enhance the learning ability of the model. After verification, the method can effectively detect prohibited item.

3 Network Architecture

The overall algorithm framework is shown in Fig. 1. The design is based on PP-YOLO, and is improved and redesigned from the following aspects:

(1) Feature Extraction Network: ResNet 50 is chosen as the feature extraction network. Compared with the large darknet53 network, ResNet network has fewer parameters, wider application scenarios and diversified branches.

(2) FPN module: FPN is feature pyramid network, which is mainly used to fuse feature maps with different resolutions generated by feature extraction network. Feature maps with larger resolution have accurate position information, while feature maps with smaller resolution have strong semantic information. Using FPN can effectively improve model performance.

(3) Detection head: The detection head consists of two convolutional layers, a 3 × 3 convolutional layer and a 1 × 1 convolutional layer. The fused features are processed by the two convolutional layers and the results are output. If there are K categories, the output dimension is 3×(K+5).

(4) MSF Module: The MSF Module further enhances the feature map generated by FPN, and the feature map with richer semantic information is produced by fusing the lower-level feature map with the higher-level feature map.

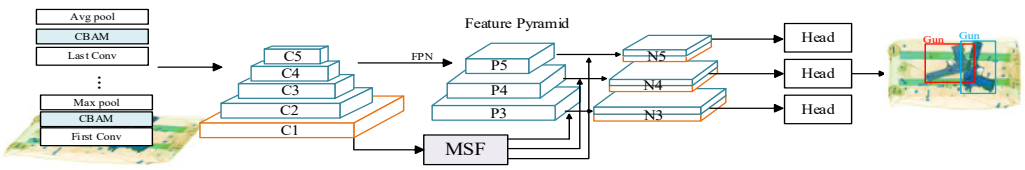


Fig. 1. Network structure image

The steps of X-ray image prohibited item detection by this framework are as follows:

Step 1. Extract the original X-ray image features by improved feature extraction network, that is, output corresponding feature maps (C1, C2, C3, C4, C5) in five stages respectively by apply ResNet 50.

Step 2. Generate feature maps (P5, P4, P3) with different resolutions through FPN network.

Step 3. Send the feature map of C1 layer to MSF module for processing, generate three feature maps with the same dimension and size as P3, P4 and P5, and fuse them (dimension splicing), and finally generate three feature maps of N3, N4 and N5.

Step 4. Use the detection head to predict the class and position information, and generate the class and accurate position information of contraband.

4. Analysis and Improvement of Network Model

The general object detection network needs to be adjusted or improved to obtain better accuracy when it is applied to a specific task. For the application of contraband recognition in X-ray images, this part elaborates the targeted model adjustment and optimization process based on PP-YOLO in detail.

4.1 Feature Extraction Network Combined with CBAM

In order to enhance the feature extraction ability of ResNet backbone network, and solve the loss problem caused by different contributions of features of different channels and spatial positions to the final recognition results, CBAM (Convolutional Block Attention Module) [35] module is introduced into the feature extraction network, which is an attention mechanism module combining spatial and channel. This module is simple and effective for feedforward CNN, and its structure is shown in Fig. 2:

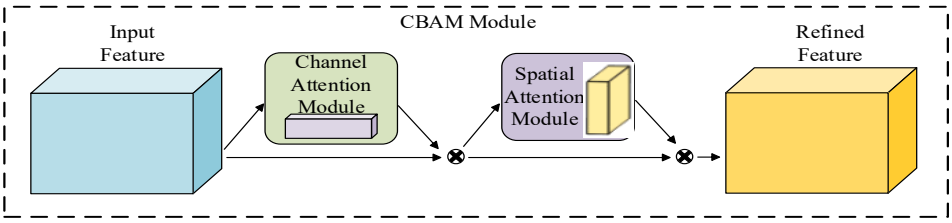


Fig. 2. The structure of CBAM module

CBAM consists of two parts, namely channel attention module and spatial attention module, and its structure is shown in Fig. 3 and Fig. 4:

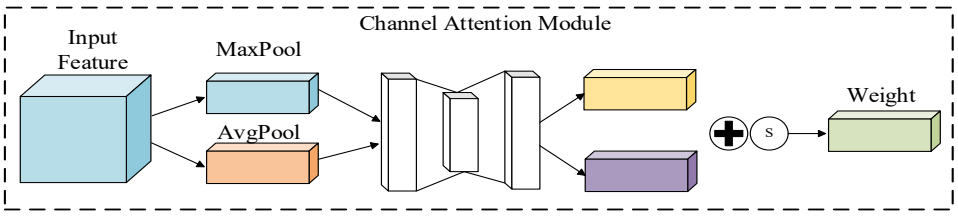


Fig. 3. Channel attention module

In the channel domain, the feature F aggregates the spatial information of the input feature through average pooling and maximum pooling, and outputs the processed maximum pooling feature F_{Max}^{c} and average pooling feature F_{Avg}^{c} by Sigmoid activation function σ. The channel attention feature $M_c(F)$ can be expressed by Equation 1:

$$\begin{aligned} M_c(F) &= \sigma(MLP(MaxPool(F)) + MLP(AvgPool(F))) \\ &= \sigma(W_1(W_0(F_{Max}^{C})) + W_1(W_0(F_{Avg}^{c}))) \end{aligned} \quad . \tag{1}$$

W_0 and W_1 are two layers in MLP. The values of W_0 and W_1 can be obtained by learning, and can be obtained $M_c(F)$ in channel domain by average pooling and maximum pooling.

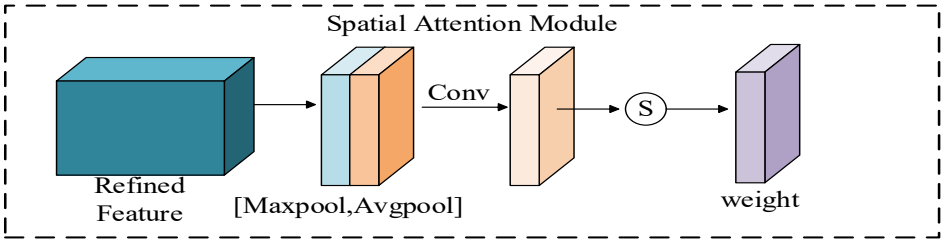


Fig. 4. Spatial attention module

In the spatial domain, the F_{Max}^{c} and F_{Avg}^{c} obtained by maximum pooling and average pooling in the channel dimension, and the obtained spatial attention feature $M_c(F)$ can be expressed by Equation 2:

$$M_s(F) = \sigma(f^{7\times7}([F_{\text{avg}}^{s}; F_{\text{max}}^{s}])) \quad . \tag{2}$$

Where σ is the Sigmoid activation function. The average pooled and maximum pooled feature maps are used with 7 × 7 convolutional kernels (Compared with 3×3 convolutional kernels, large convolutional kernel has better effect), the calculated feature map is activated by Sigmoid function, different weights are generated at different spatial positions, and the weights are multiplied by the feature map after passing by channel attention at corresponding positions to generate the feature map containing both channel attention and spatial attention.

Two CBAM modules are introduced into the feature extraction network, which are after the first convolutional layer before the maximum pooling layer and before the last pooling layer. Because this design would not change the structure of block in ResNet, that is, it would not change the structure of ResNet network, and it can use pre-training parameters without training from scratch, which can speed up the convergence of the model.

4.2 Network Model Compression with CompConv

The backbone network is composed of five residual blocks, which contains more convolutional layers, and the calculation of convolutional layers is time consuming. In order to detect prohibited item in real time, reduce the calculation amount of the model and speed up the inference, CompConv [36] is used to upgrade some common convolutional blocks in backbone, which can improve the running speed of the network without losing accuracy. CompConv uses divide-and-conquer method to simplify the generation of feature map, and integrates the input feature map into the output to effectively inherit the input information. Its structure is shown in Fig. 5.

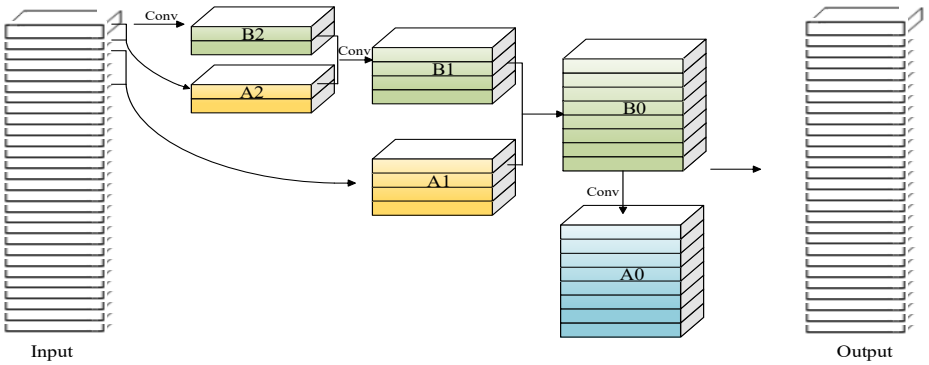


Fig. 5. CompConv core unit

To generate a feature map X of C channel, we can choose to generate two feature maps X_A and X_B , both of X_A and X_B have a number of channels of C/2, and then combine the two together by the following equation 3:

$$X = X_A + WX_B \ . \tag{3}$$

which "+" means splicing along the dimension direction, and W is a learnable parameter. Equation 3 embodies the core idea of CompConv, which means that it is completely mapped from a subset of input features and can inherit part of the original input information, and X_B represents features transformed from input features by convolutional blocks. According to Equation 3, the re-operation of X_B can be subdivided into two parts, as shown in Equation 4:

$$X_{B_i} = X_{A_{i+1}} + W_{i+1} X_{B_{i+1}} \ (\mathrm{i} = 0, \ldots, \mathrm{d} - 1,) \ . \tag{4}$$

Where d is recursive depth, A0 is treated differently as shown in Fig. 5 because there are many channels. If some channels are directly copied as A0, a lot of redundant information will be generated, so A0 is transformed from B0 by convolution.

CompConv carries out feature mapping based on divide-and-conquer method, so how to divide channels affects computational efficiency and learning ability. C_{in} and C_{out} are used to represent the number of input and output channels respectively, and C_{prim} is the smallest computing unit (such as X_{B2}) when d = 3 in Fig. 5, so the relationship between output C_{out} and C_{prim} can be described by Equation 5:

$$C_{\mathrm{out}} = \sum_{i=1}^{d} 2^i C_{prim} \ . \tag{5}$$

It can be deduced from Equation 6:

$$C_{\mathrm{prim}} = \left\lceil \frac{C_{out}}{2 \times (2^{\mathrm{d}} - 1)} \right\rceil . \tag{6}$$

It can be seen from Equation 6 that C_{prim} is highly dependent on recursive depth d, which is also a super parameter of CompConv. Larger d corresponds to higher compression ratio, and the selection of recursive depth d can be described by Equation 7:

$$d = \max(\log_2(\max(1, \frac{C_{in}}{C_0})) + 1, 3) \quad . \tag{7}$$

Where C_0 is the model-specific design choice, which is determined by the target compression ratio and the model size. In ResNet, $C_0 = 128$ is selected, assuming that both the input and output resolutions are H × W, and the computational complexity of ordinary convolution and CompConv is expressed by Equation 8 and 9, respectively:

$$O_{\text{conv}} = H \times W \times K^2 \times C_{\text{in}} \times C_{out} \quad . \tag{8}$$

$$O_{CompConv} = H \times W \times K^2 \times (C_{\text{in}} \times C_{prim} + \sum_{i=1}^{d-1} (2^i C_{prim})^2 + 2^{d-1} C_{prim}) \quad . \tag{9}$$

Under the configuration of $C_{in} = C_{out}$ and d = 3, according to Equations 8 and 9, CompConv can generate feature maps with the same output dimension with only 20% computing resources, which reduces the computation of the network.

The position of CompConv in the ResNet network is shown in Fig. 6.

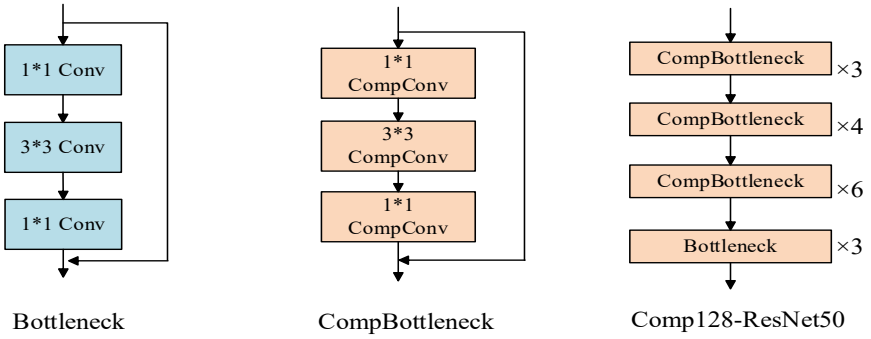


Fig. 6. CompConv bottleneck structure image

4.3 Feature Fusion with Accurate Position Information

Although FPN network fuses feature maps with different resolutions, and enriches the information of the feature map, small instances are easy to be lost in the process of convolution. In order to enhance the detection ability of the model for small instances, multi-scale fusion module (MSF) is added to the network, and the feature map of C1 layer is converted into the feature map with the same shape as P3, P4 and P5 by several convolutional layers and pooling layers, and the more accurate position information and semantic information are fused. The specific parameters are shown in Table 1.

Table 1. Network input and output table

Input	Output				
	C1	C2	C3	C4	C5
320	160	80	40	20	10
416	208	104	52	26	13
512	256	128	64	32	16
608	304	152	76	38	19

According to the table, when the size of the input shape is (W, H), the output size of Ci layer is $(\frac{W}{2^i}, \frac{H}{2^i})$. After a stage, the backbone network reduces the size of the feature map by half, but the dimension of the feature map doubles in order to retain as much image information as possible while reducing the size. According to the data of this table, the MSF module is designed, and the specific structure is shown in Fig. 7.

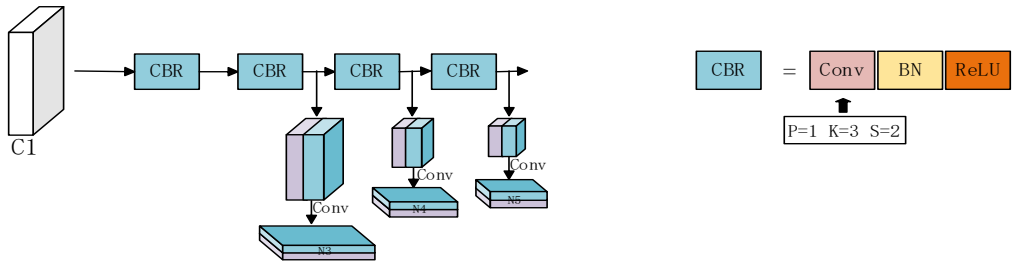


Fig. 7. MSF structure image

The input of MSF module is the output of the first stage of ResNet, and its basic module is CBR (Conv BN ReLU), which is composed of convolutional layer, normalization layer and activation function. The convolutional layer in the module acts as down sampling, and the input and output sizes of the network are shown in Equation 10:

$$\text{Out} = (In - F + 2P)/S + 1 \ . \tag{10}$$

Where Out represents the size of the output feature map, In represents the input size, F is the convolutional kernel size, P is the filled pixel value, S represents the step size, 3*3 convolutional kernel is used, and the same filling method is adopted to ensure that the output feature map size is the same as the input feature map size.

After the first CBR processing, the output feature map size is $(\frac{W}{2^2}, \frac{H}{2^2}, 256)$, and the output after the second CBR is $(\frac{W}{2^3}, \frac{H}{2^3}, 512)$. The output size is the same as the size and dimension of P3. The output of the second CBR is spliced with P3, and the splicing process can be described by Equation 11:

$$Z_{concat} = \sum_{i=1}^{c} X_i * K_i + \sum_{i=1}^{c} Y_i * K_{i+c} \quad . \tag{11}$$

Where Z_{concat} represents the number of output channels, X_i is the channel of the first feature map, Y_i is the channel of the second feature map, K_i is the convolutional kernel, and then the channel is adjusted by convolutional layer to finally generate N3 feature map. The generation of N4 and N5 feature maps is similar to N3, and the generated N3, N4 and N5 feature maps are detected to produce predicted results.

5. Experimental Dataset and Data Preprocessing

5.1 Introduction to Datasets

At present, there are few datasets for prohibited items detection, and two common datasets are GDXray and SIXray. Among them, GDXray dataset contains 19,047 X-ray images, and includes five parts, such as castings,

welds, luggage, natural objects and background. The dataset sample is shown in Fig. 8.

Fig. 8. Random sample of GDXray dataset

The images in GDXray dataset are relatively simple, and most of them are X-ray images taken by single-energy security detection machine under ideal conditions. In the real scene, X-ray images often overlap and have various angles, so dual-energy security detection instruments are more commonly used in security detection places at present. Therefore, in order to verify the validity of the model, we mainly chooses SIXray dataset as experimental data.

SIXray dataset contains 1,059,231 X-ray images, of which 8,929 images contain contraband, 3,131 images contain guns (class 1), 3,961 images contain pliers (class 2), 2,199 images contain wrenches (class 3), 1,943 images contain knives (class 4), 983 images contain scissors (class 5), and only 60 images contain hammers.

Therefore, only the first five categories of contraband are identified. Some datasets are shown in Fig. 9.

Fig. 9. SIXray database partial sample

5.2 Data Preprocessing

For the original dataset, we done the following processing: First of all, the tag files are counted, and the results show that most of the marked objects belong to large and medium objects. There are two ways to increase the amount of prohibited item in the dataset. The first is the traditional data enhancement method, which directly scales, rotates and transforms the original image, which can increase the amount of data. In order to better reflect the overlapping characteristics of X-ray contraband, we also uses the second data enhancement method.

Image Acquisition and Preprocessing of Single Prohibited Item. A single contraband image is obtained by manual segmentation, as shown in Fig. 10.

Fig. 10. The single prohibited item image

Rough single contraband image can be obtained by manual segmentation. The reason for the roughness is there are burrs at the edge of the obtained single contraband, so morphological methods are needed for data processing. The single contraband image is corroded, and the refined contraband image is shown in Fig. 11.

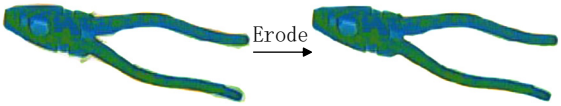


Fig. 11. Contraband image after etching operation

Etching operation can make the image shrink inward along the boundary. After etching, it can be clearly seen that the boundary of the image has been optimized to a certain extent, and the contour information can be displayed more clearly.

For the pictures after corrosion operation, it is necessary to mark the data to simplify the generation of datasets. There are six categories of prohibited item, and LabelImg software is used to mark the data and generate the corresponding marking files. In order to reflect the random placement of contraband, in this paper, the pictures and annotation files of contraband are processed by affine transformation, including random rotation, scaling, 3D rotation and image flip, which provides diversified data support for the next image fusion. The enhanced single contraband image is shown in Fig. 12.

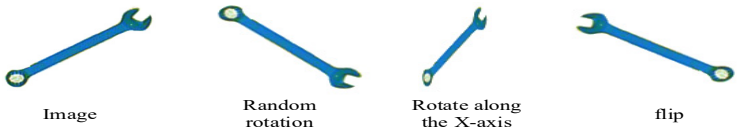


Fig. 12. Image after affine transformation

Affine transformations can be described by Equation 12:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} R_{00} & R_{01} & T_x \\ R_{10} & R_{11} & T_y \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{12}$$

Where T_x and T_y represents the amount of translation, and x and y represent the coordinate points before transformation. x' and y' represent the transformed coordinates, and the parameter R is the information such as rotation and scaling of the image.

At the same time of image transformation, the corresponding json file should be processed, and the tag points in it should be corrected to the transformed coordinate information. After affine transformation, the data preparation is completed. In the next step, the prepared data and the real X-ray image background will be fused to generate the X-ray contraband enhancement dataset with tag information.

Image Fusion. Image fusion is divided into the following steps:

Step 1: The background of the picture is transparent. Calculate each pixel of the picture respectively. After experimental verification, it is better to take the threshold value of 200 for processing. Set the value of RGB three channels of pixels with a threshold value greater than 200 to (255, 255, 255), and set the transparency to 0, that is, full transparency.

Step 2: Randomly select the background picture and the transformed contraband picture, and calculate the random coordinate points as the upper left corner coordinates of the small picture. In order to prevent the small picture from crossing the boundary in the final generated picture, the range of coordinates is restricted. The value range of x is (0, W_{large}-W_{small}), the range of y is (0, H_{large}-H_{small}), where W_{large} and H_{large} are the width and height of the background image respectively, and W_{small} and H_{large} are the width and height of the contraband image. After obtaining the upper left corner coordinates, the small image is fused into the large image, and the information of the marking file is modified. In order to reflect the penetration characteristics of the X-ray image, we set the transparency of small images to 0.82 during the fusion process. In order to reflect the overlapping characteristics, a plurality of contraband images are pasted in one image, and there is a 50% probability that two contraband images will overlap. After the processing of the above steps, the dataset of marked files can be generated without marking one by one. After the processing of these two methods, the amount of contraband in the dataset is greatly increased, and finally 51034 X-ray images are generated. In this paper, we divide the data set into training set, verification set and detection set according to the ratio of 7: 2: 1. Some data are shown in Fig. 13.

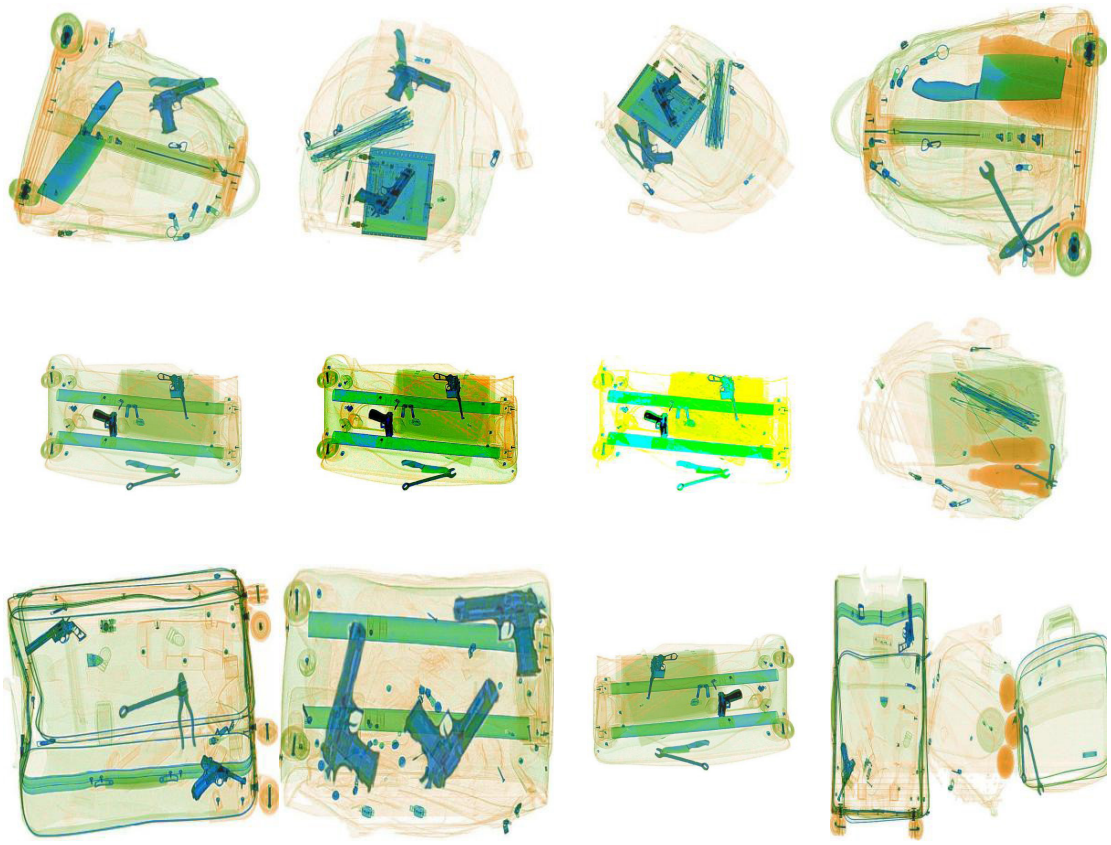


Fig. 13. Partial dataset image

6 Analysis of Experimental Results

The experimental environment and hardware conditions: The hardware environment is 3.0 GHz CPU, 16G memory, NVIDIA GeForce 1080Ti graphics card × 2, 11G video memory, and the software environment is Ubuntu 20.04 operating system, using Python language and PaddlePaddle deep learning framework.

6.1 Experimental Design

In order to verify the effectiveness of the improved method for prohibited item detection, the following experiments are carried out.

Experiment 1: Compared the native PP-YOLO network with YOLOv3, YOLOv4 and YOLOv5.
Experiment 2: Integrate CBAM module into PP-YOLO network.
Experiment 3: Add MSF to FPN network.
Experiment 4: Add experiment of CC to ResNet.
Experiment 5: Add CBAM and MSF to PP-YOLO.
Experiment 6: Add CBAM and CC to ResNet.
Experiment 7: MSF and CC join PP-YOLO network.
Experiment 8: Add CBAM, MSF and CC to PP-YOLO network.

6.2 Analysis of Results

The single-stage detector with better detection performance is selected for comparative experiments, and the experimental results are shown in Table 2.

Table 2. Identification results of experiment (%)

Model	AP for Gun (class1)	AP for Knife (class2)	AP for Pliers (class3)	AP for Wrench (class4)	AP for Scissors (class5)	mAP
YOLOv5	96.10	91.70	**95.0**	93.20	91.8	**93.56**
YOLOv4	**96.60**	91.0	94.4	93.31	**92.1**	93.48
PP-YOLO [37]	95.80	**92.2**	94.9	**94.32**	89.7	93.38
YOLOv3	88.04	76.61	56.24	61.02	45.3	65.44

It can be seen from the table that all models except YOLOv3 have high recognition accuracy for class 1 and class 3. There is little difference between class 1 and class 3 in the dataset and there are more training data, so the recognition accuracy is high. YOLOv3 appeared earlier and did not use too many data enhancement methods, so the detection accuracy is lower than other detectors, and the recognition effect for class 5 is the worst. First, class 5 occupies a small area in the picture, and in many cases it is in an occluded state, which leads to low recognition accuracy. The experimental results of YOLOv5, YOLOv4 and PP-YOLO networks on this dataset are not much different, and the recognition accuracy is better than that of YOLOv3. On the whole, YOLOv5 network has achieved the best detection effect, while the accuracy of PP-YOLO detector is slightly lower than that of YOLOv5.

YOLO series is a typical single-stage detector, and RCNN series networks all have RPN stages, which produce suggestions on the possible areas of targets. Faster-RCNN works well in double-stage detectors, so experiments are carried out on Faster-RCNN networks, and the experimental results are shown in Table 3.

Table 3. Comparison of double-stage and single-stage network experiments (%)

Model	AP for Gun	AP for Knife	AP for Pliers	AP for Wrench	AP for Scissors	mAP	FPS
Faster-RCNN [17]	90.21	86.36	90.21	87.32	**91.63**	89.14	22
PP-YOLO	**95.80**	**92.2**	**94.9**	**94.32**	89.7	**93.38**	**47**
SSD [23]	83.32	82.63	78.45	79.85	86.21	82.09	37

It can be seen from Table 3 that in this dataset, the detection accuracy of Faster-RCNN is better than SSD, but SSD has great advantages in speed, and PP-YOLO network has great advantages in processing speed and accuracy.

In order to verify the module proposed to improve the performance of PP-YOLO, the following Ablation experiments are done, the experimental results are shown in Table 4.

Table 4. Ablation experiments on PP-YOLO network (%)

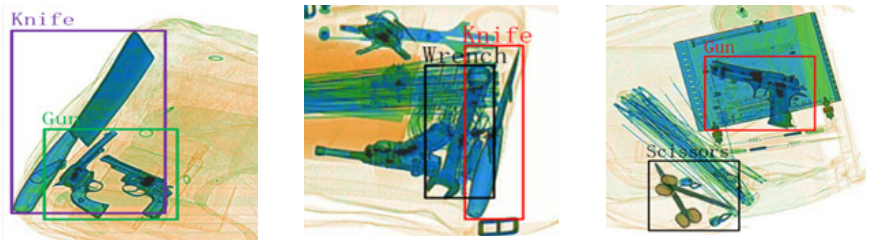
Model	CBAM	MSF	CC	AP for Gun	AP for Knife	AP for Pliers	AP for Wrench	AP for Scissors	mAP	FPS
Base				95.80	92.20	94.9	94.32	89.70	93.38	47
2	√			96.04	92.34	95.12	94.51	89.93	93.59	45
3		√		96.31	92.62	95.33	94.68	93.21	94.43	40
4			√	95.91	92.22	94.81	94.20	89.64	93.36	**55**
5	√	√		**96.61**	**93.03**	95.43	**94.96**	93.34	**94.67**	39
6	√		√	96.10	92.38	95.03	94.58	90.05	93.62	52
7		√	√	96.25	92.67	95.25	94.71	93.15	94.41	46
8	√	√	√	96.59	93.00	**95.48**	94.85	**93.43**	**94.67**	45

Experiments are carried out on the original network, and the results are shown in the first row of Table 4. Except for the poor recognition accuracy of class 5, the recognition accuracy of other categories is over 90%. By introducing CBAM module, the paper assigns weights to different channels and spatial positions, which improves mAP by 0.21%. The introduction of MSF module improves mAP by 1 percentage point, especially for class 5 recognition accuracy, reaching 3%, which also shows that the lower network layer has more accurate position information, and more information can be retained to the greatest extent after a small number of convolutional layers. CompConv operation has little influence on the accuracy, but it improves the final recognition speed and basically eliminates the computation problem caused by CBAM and MSF. By combining CBAM and MSF modules, the recognition accuracy of this network is relatively the best, reaching 94.67%, which is 1.11% higher than YOLOv5. Experimental results show that the proposed MSF module and CBAM module can effectively improve the detection accuracy of the model on SIXray datasets, and can also obtain faster detection speed under the action of CompConv module.

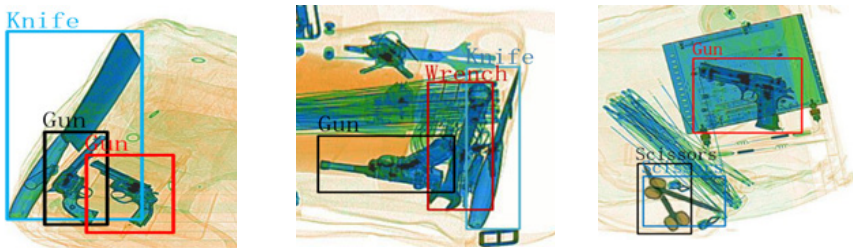
6.3 Visualization of Detection Results

There are some problems in X-ray images, such as overlapping of some objects, small number and small area, which lead to some missing detection phenomena in the original network. As shown in the first column of Fig. 14, the scissors in the original network are not completely detected because their metal parts are small and some of them are blocked by foreground objects. By adding CBAM module, the problem of missing detection caused by overlap can be effectively solved, and MSF module can transmit the accurate position information of lower layer. Compared with the detection effect only under adding CBAM module, a more accurate position box is produced under the joint action of CBAM and MSF, as shown in the third column of Fig. 14.

(1) Baseline



(2) +CBAM



(3) +CBAM & MSF

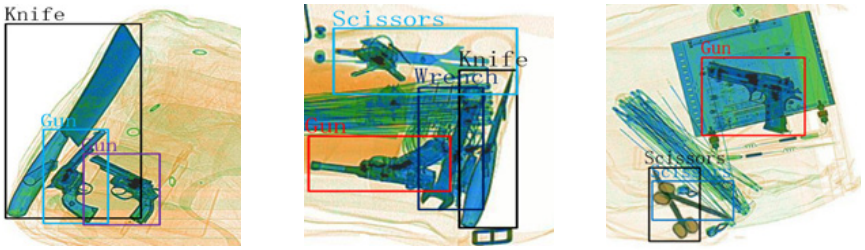


Fig. 14. Detection effect image under the action of different modules

6.4 Comparison between This Method and Other Methods

In order to verify the effectiveness of the algorithm in this paper, the performance of the network in this paper is compared with that of other networks. Due to different data sets, the results are slightly different, but the overall segmentation can still be used as the basis for comparison. The comparison results are shown in Table 5. It can be seen from the table that compared with two-stage detectors, the network performance in this paper is better and the average accuracy is improved. The release time of YOLOv5 network is later than that of PP-YOLO network, and a large number of tricks are integrated. The average accuracy is higher than that of PP-YOLO network. However, when the imaging angle of the target object changes greatly, or the lines in the background are more complex, there is still missing detection.

Table 5. Performance comparison of different algorithms (%)

Model	AP for Pliers	AP for Scissors	AP for Knife	AP for Gun	AP for Wrench	mAP	Recall	FPS
YOLO-C	60.18	58.14	83.12	93.11	73.82	73.68	-	40
MFFNet [38]	82.22	72.24	75.29	90.42	71.17	78.27	-	19
YOLOv3 [39]	84.04	80.34	75.67	93.88	87.18	86.59	-	7.8
ASPP-YOLOv4 [40]	87.36	83.76	81.39	95.78	77.84	85.23	75.16	-
Res152-YOLO [41]	96.1	97.3	91.5	98.1	92.8	95.16	67.12	40
YOLOv5s-AFA [42]	-	-	-	-	-	95.6	88.9	-
FEFNet [43]	85.95	84.0	81.43	95.15	81.65	85.64	-	31.57
Faster R-CNN [44]	81.96	79.11	76.02	95.53	68.19	80.16	-	25.86
ACMNet	85.9	80.3	80.2	91.5	83.6	84.3	-	-
EM2Det [45]	89	84	79	98	77	85.4	-	8
CenterNet [46]	89.9	77.43	91.88	96.4	85.9	88.3	-	-
Ours	95.48	93.43	93.00	96.59	94.85	94.67	83.8	45

7 Conclusion

This paper proposes a threat detection network based on PP-YOLO. By adding MSF module to fuse the features of lower layers, the detection accuracy of small objects (Scissors) increases by 3.51%, indicating that the lower layer has more location information, and the fusion of the information of lower layers can effectively increase

the detection accuracy. CBAM attention mechanism is added to the backbone network, which makes the weights generated by the network at different locations of different feature maps different, makes the network pay more attention to the channels and locations where there are targets, enhances the feature extraction capability of the model, and can produce more accurate boundary boxes. The CompConv module is added. Compared with the original network, the processing speed is increased by 7 FPS. Finally, the experiment was carried out on the data set, and the average recognition accuracy reached 94.67%, and the detection speed reached 45FPS, which can detect contraband to a certain extent.

Although the attention mechanism can increase the detection accuracy of the model, the detection effect of highly overlapping contraband is still poor, which is likely to lead to missed detection. In addition, the detection accuracy of small target contraband needs to be improved. Therefore, the future research direction should focus on highly overlapping contraband detection and small target contraband detection.

Acknowledgement

The authors express their acknowledgement for the anonymous review.

References

[1] D. Turcsany, A. Mouton, T.-P. Breckon, Improving feature-based object recognition for X-ray baggage security screening using primed visualwords, in: Proc. IEEE International Conference on Industrial Technology, 2013.
[2] Y. Wei, Z. Zhu, H. Yu, W. Zhang, An automated detection model of threat objects for X-ray baggage inspection based on depthwise separable convolution, Journal of Real-Time Image Processing 18(3)(2021) 923-935.
[3] D. Mery, Inspection of Complex Objects Using Multiple-X-Ray Views, IEEE/ASME Transactions on Mechatronics 20(1)(2015) 338-347.
[4] D. Mery, V. Riffo, I. Zuccar, C. Pieringer, Object recognition in X-ray testing using an efficient search algorithm in multiple views, Insight-Non-Destructive Testing and Condition Monitoring 59(2)(2017) 85-92.
[5] Q. Wang, K. Wu, X. Wang, Y. Sun, X. Yang, X. Lou, Automatic detection and classification of foreign bodies of dumplings based on x-ray, Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/Journal of Computer-Aided Design and Computer Graphics 30(12)(2018) 2242-2252.
[6] J.-Y. Liu, J. Leng, Y. Liu, Deep Convolutional Neural Network Based Object Detector for X-Ray Baggage Security Imagery, in: Proc. 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), 2019.
[7] R.-L. Galvez, E.-P. Dadios, A.-A. Bandala, R. Vicerra, YOLO-based Threat Object Detection in X-ray Images, in: Proc. 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), 2019.
[8] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[9] X.-L. Ji, J. Wu, Y. J.-B. Yi, X.-G. Zhang, Automatic detection algorithm for controlled items based on deep learning, Laser & Optoelectronics Progress 56(18)(2019) 180402.
[10] Y.-T. Zhang, H.-G. Zhang, T.-F. Zhao, J.-F. Yang, Automatic detection of prohibited items with small size in x-ray images, Optoelectronics Letters 16(4)(2020) 313-317.
[11] S.-X. Guo, L. Zhang, Yolo-C: One-Stage Network for Prohibited Items Detection Within X-Ray Images, Laser & Optoelectronics Progress 58(8)(2021) 67-76.
[12] M. Chouai, M. Merah, M. Mimi, Ch-net: deep adversarial autoencoders for semantic segmentation in x-ray images of cabin baggage screening at airports, Journal of Transportation Security 13(1-2)(2020) 71-89.
[13] Z.-G. Su, S.-Q. Yao, A Multi-object Prohibited Items Identification Algorithm Based on Semantic Segmentation, Journal of Signal Processing 36(11)(2020) 1940-1946.
[14] Y. Zhang, W. Kong, D. Li, X. Liu, On Using XMC R-CNN Model for Contraband Detection within X-Ray Baggage Security Images, Mathematical Problems in Engineering 2020(2020) 1-14.
[15] T. Hassan, M. Bettayeb, S. Akay, S. Khan, M. Bennamoun, N. Werghi, Detecting Prohibited Items in X-Ray Images: a Contour Proposal Learning Approach, in: Proc. IEEE Conference on Image Processing, 2020.
[16] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[17] R. Girshick, Fast R-CNN, in: Proc. of the IEEE international conference on computer vision, 2015.
[18] S.-Q. Ren, K.-M. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transactions on Pattern Analysis & Machine Intelligence 39(6)(2017) 1137-1149.
[19] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proc. 2016

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[20] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[21] J. Redmon, A. Farhadi, YOLOv3: An Incremental Improvement, arXiv e-prints. <https://arxiv.org/abs/1804.02767>, 2018.
[22] A. Bochkovskiy, C.-Y. Wang, H. Liao, Yolov4: optimal speed and accuracy of object detection, arXiv preprint. <https://arxiv.org/abs/2004.10934>, 2020.
[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.-C. Berg, in: Proc. SSD: single shot multibox detector, 2016.
[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, IEEE Transactions on Pattern Analysis & Machine Intelligence 42(2)(2020) 318-327.
[25] Y.-K. Zhang, Z.-G. Su, H.-G. Zhang, J.-F. Yang, Multi-scale prohibited item detection in X-ray security image, Journal of Signal Processing 36(7)(2020) 1096-1106.
[26] C. Zhu, B.-Y. Li, X.-Q. Liu, Z.-N. Feng, A deep convolutional neural network based on YOLO contraband detection, Journal of Hefei University of Technology (Natural Science) 44(9)(2021) 1198-1203.
[27] S.Q. Mu, J.J. Lin, H.Q. Wang, X.Z. Wei, An Algorithm for Detection of Prohibited Items in X-ray Images Based on Improved YOLOv4, Acta Armamentarii 42(12)(2021) 2675-2683.
[28] D. Qi, X. Jin, H. Wu, Application of Multi-Scale Fractal Feature in Defects Detection of Log X-Ray Image, in: Proc. 2009 WRI Global Congress on Intelligent Systems, 2009.
[29] S.-Q. Yao, Z.-G. Su, Prohibited Item Identification Algorithm Based on Lightweight Segmentation Network, Laser And Optoelectronics Progress 58(2)(2021) 219-227.
[30] L.-R. Li, P. Chen, Y.-L. Zhang, Insulator defect Detection based on multi-scale feature coding and dual attention fusion, Laser And Optoelectronics Progress 59(24)(2022) 338-348.
[31] H.-T. Cao, H.-J. Shi, X.-L. Song, M.-J. Li, H.-L. Dai, Z. Huang, Prediction of Pedestrian Intention and Trajectory Based on Multi-feature Fusion, China Journal of Highway and Transport 35(10)(2022) 308-318.
[32] L.-N. Sun, P.-X. Yuan, Research of detection technology in X-ray security inspection Equipment, China Measurement Technology 32(3)(2006) 20-22, 121.
[33] Y.-S. Lu, Y.-J. Tang, X.-R. Ma, Low contrast filament sizing defect detection method of non-woven fabric based on deep feature fusion, Journal of Computer Applications 42(5)(2022) 1440-1446.
[34] L. Zhang, Y.-G. Shan, J. Yuan, Target tracking based on conditional confrontation network and hierarchical feature fusion, Computer Engineering and Applications 58(23)(2022) 221-229.
[35] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: Convolutional Block Attention Module, in: Proc. European Conference on Computer Vision, 2018.
[36] C. Zhang, Y. Xu, Y. Shen, CompConv: A Compact Convolution Module for Efficient Feature Learning, in: Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021.
[37] X. Long, K. Deng, G. Wang, Y. Zhang, S. Wen, PP-YOLO: An Effective and Efficient Implementation of Object Detector, 2020.
[38] Y.-X. Wang, L. Zhang, Dangerous Goods Detection Based on Multi-Scale Feature Fusion in Security Images, Laser & Optoelectronics Progress 58(8)(2021) 144-151.
[39] C. Zhu, B.-Y. Li, X.-Q. Liu, Z.-N. Feng, A deep convolutional neural network based on YOLO for contraband detection, Journal of Hefei University of Technology (Natural Science) 44(9)(2021) 1198-1203.
[40] H.-B. Wu, X.-Y. Wei, M.-H. Liu, A.-L. Wang, H. Liu, J.-Y. Iwahori, Improved YOLOv4 for dangerous goods detection in X-ray inspection combined with atrous convolution and transfer learning, Chinese Optics 14(6)(2021) 1417-1425.
[41] J.-C. Yang, J.-H. Huang, Y.-L. Han, P. Wang, X.-H. Li, X-Ray Security Image Detection Network Based on Optimized YOLOv4, Computer Systems & Applications 30(12)(2021) 116-122.
[42] H. Zhang, S.-C. Zhang, Security Inspection Image Object Detection Method with Attention Mechanism and Multilayer Feature Fusion Strategy, Laser & Optoelectronics Progress 59(16)(2022) 187-198.
[43] C. Li, H. Zhang, Z.-Q. Zhang, A.-B. Che, Y.-N. Wang, Integrated multi-scale features and global context in X-ray detection for prohibited items, Journal of Image and Graphics 27(10)(2022) 3043-3057.
[44] J.-N. Kang, L. Zhang, Multi-scale X-Ray Security Inspection Image Detection with Multi-channel Region Proposal, Computer Engineering and Applications 58(1)(2022) 224-231.
[45] K. Zhang, L. Zhang, Multi-Scale Detection for X-Ray Prohibited Items in Complex Background, Laser & Optoelectronics Progress 58(22)(2021) 102-112.
[46] J.-Q. Qiao, L. Zhang, X-Ray Object Detection Based on Pyramid Convolution and Strip Pooling, Laser & Optoelectronics Progress 59(4)(2022) 209-220.