Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6544325, 14 pages
https://doi.org/10.1155/2021/6544325

Research Article

X-Ray Image Recognition Based on Improved Mask R-CNN Algorithm

Jicun Zhang,[1,2] **Xueping Song**,[1] **Jiawei Feng**,[2] **and Jiyou Fei**[1]

[1]*School of Mechanical Engineering, Dalian Jiaotong University, Dalian, Liaoning 116028, China*
[2]*Neusoft Group (Dalian) Co., Ltd., Dalian, Liaoning 116085, China*

Correspondence should be addressed to Jicun Zhang; zhangjicun89@163.com and Jiyou Fei; fjy@djtu.edu.cn

Received 1 August 2021; Accepted 21 August 2021; Published 7 September 2021

Academic Editor: Shanglei Jiang

It is an important part of security inspection to carry out security and safety screening with X-ray scanners. Computer vision plays an important role in detection, recognition, and location analysis in intelligent manufacturing. The object detection algorithm is an important part of the intelligent X-ray machine. Existing threat object detection algorithms in X-ray images have low detection precision and are prone to missed and false detection. In order to increase the precision, a new improved Mask R-CNN algorithm is proposed in this paper. In the feature extraction network, an enhancement path is added to fuse the features of the lower layer into the higher layer, which reduces the loss of feature information. By adding an edge detection module, the training effect of the sample model can be improved without accurate labeling. The distance, overlap rate, and scale difference between objects and region proposals are solved using DIoU to improve the stability of the region proposal's regression, thus improving the accuracy of object detection; SoftNMS algorithm is used to overcome the problem of missed detection when the objects to be detected overlap each other. The experimental results indicate that the mean Average Precision (mAP) of the improved algorithm is 9.32% higher than that of the Mask R-CNN algorithm, especially for knife and portable batteries, which are small in size, simple in shape, and easy to be mistakenly detected, and the Average Precision (AP) is increased by 13.41% and 15.92%, respectively. The results of the study have important implications for the practical application of threat object detection in X-ray images.

1. Introduction

At present, in the express logistics industry, high-speed railway stations, airports, and other public transportation fields, the security detection of parcels, baggage, and other items is mainly realized by X-ray security machines plus manual detection by human inspectors. The detection results mainly rely on the experience of screeners, which has high input cost and low efficiency. There are many kinds of items to be detected, which may be likely to be covered by other items, so the false detection and missed detection of manual detection occur from time to time.

Computer vision plays an important role in detection, recognition, and location analysis in intelligent manufacturing. With the development of artificial intelligence, automatic detection can be realized with the help of computer vision technology to replace manual detection. The object detection algorithm is an important part of computer vision.

Recently, deep learning has shown promising results in many image-based tasks. Convolutional neural networks (CNNs) [1] are the derivatives of deep learning, which have been widely used in various applications, such as medical image analysis and applications [2–4], face detection [5], speech recognition [6], pose estimation [7], and other computer vision tasks.

The image recognition technology is integrated into the X-ray security scanning machine to realize the automatic detection of threat items, which can greatly reduce the workload of human inspectors and improve detection efficiency and precision. It has important significance in the field of security inspection of intelligent logistics and intelligent transportation.

The research of threat object detection in X-ray images has also made some progress [8–10]. In order to generate a large number of X-ray images of bags with threat objects, Threat Image Projection (TIP) [11] was developed. A real-time TIP model based on deep learning was proposed. Therefore, this model can be used to train screeners to recognize threat objects in real-time TIP images and can be applied to automated detection of threat objects research in the future.

Mery et al. [12] attempt to make a contribution to the field of object recognition in X-ray testing by evaluating different computer vision strategies that have been proposed in the last years, such as BoWs, sparse representations, deep learning, and classic pattern recognition schemes, among others. The author believes that a CNN trained with X-ray images (instead of optical images) would lead to better results in X-ray testing.

Akcay et al. [10] compare several object detection methods and come to a conclusion: it shows that contemporary Faster R-CNN, R-FCN, and YOLOv2 approaches outperform SW-CNN, which is already empirically shown to outperform handcrafted features regarding both speed and accuracy. CNN features achieve superior performance to handcrafted BoVW features.

Gao et al. [13] propose the combination of Faster R-CNN algorithm and Feature Pyramid Network to realize the detection of small items on clothes, which shows that Faster R-CNN is effective for small items detection.

Gaus et al. [14] propose a dual CNN architecture to firstly isolate liquid and electrical objects by type and subsequently screen them for abnormalities.

The abovementioned convolutional neural networks have made some achievements in the field of threat object detection in an X-ray image, but they cannot reach the practical application level for the safe detection of Real-time X-ray images. The main difficulties in X-ray detection of threat objects are as follows: (1) Objects are blocked and overlapped, which are easy to be undetected. (2) Some threat objects are small in size and simple in shape, such as knives and portable batteries, which are easily confused with other items in X-ray images, and the object detection precision is low.

There are two kinds of object detection algorithms based on the neural network: one-stage detection algorithm and two-stage algorithm. One-stage detection algorithm is represented by YOLO [15] and SSD [16]. The full name of YOLO is "You Only Look Once," which means that the algorithm only needs one CNN operation. YOLO uses an end-to-end unified, fully convolutional network structure that predicts the objectless assurance and the bounding boxes concurrently over the whole image. YOLO resizes the input image, runs a single convolutional network on the image, and thresholds the resulting detections by model's confidence.

The full name of SSD algorithm is "single shot multibox detector." Single shot indicates that SSD algorithm belongs to the one-stage method, and multibox indicates that SSD is multibox prediction. It combined the anchor mechanism in the Faster R-CNN and the regression idea in YOLO, as the input image feature extraction using a small convolution filter, and the feature of the different scales with different aspect ratio classification prediction.

The one-stage algorithm directly extracts features from the network to predict object classification and location, the speed is fast, but the accuracy is not as high as that of the two-stage algorithm. It is widely used in the field of video stream object detection with high real-time requirements [17–21].

The two-stage detection algorithm is represented by the R-CNN algorithm. The two-stage algorithm needs to generate a region proposal (a preselected box that may contain the object to be inspected) and then classify each candidate box (the position will also be corrected). This kind of algorithm is relatively slow because it needs to run the detection and classification process many times, but the accuracy is high.

The application scene of X-ray image threat object detection requires high recognition accuracy. So, the R-CNN algorithm with two-stage detection is more suitable for this scene. In this paper, the Mask R-CNN algorithm, which has a good effect in the field of object detection, is selected and optimized. The optimized algorithm has improved the mean Average Precision (mAP) more than the original Mask R-CNN algorithm and also improved a lot compared with other algorithms, which has the practical value of threat object detection in X-ray images.

The main contributions of this paper are as follows:

(1) An improved algorithm based on Mask R-CNN is proposed. By optimizing in the network layer, the loss of feature information is reduced; online hard negative example mining (OHEM) is used to improve the robustness of the model.

(2) Using DIoU instead of IoU makes the object region overlap more with the region proposal and the regression effect is better; SoftNMS algorithm is used to replace the NMS algorithm, which increases the object detection rate in the overlapping area of threat objects.

(3) We propose an X-ray image dataset for model training and testing. Our experimental results show that the performance of the proposed method outperforms the state-of-the-art object detection method in precision and recall rate, especially for overlapped objects and threat objects with small sizes and simple shapes.

2. Materials and Methods

The development process of the R-CNN algorithm is shown in Figure 1, from R-CNN [22], Fast R-CNN [23], Faster R-CNN [24] to the most advanced Mask R-CNN [25] algorithm at present, which is manifested in the continuous improvement of precision and speed, covering various fields from classification to detection, segmentation and positioning [26–30]. The two-stage detection has high detection precision and strong robustness, but the detection speed is slow. The one-stage detection has a simple structure and fast

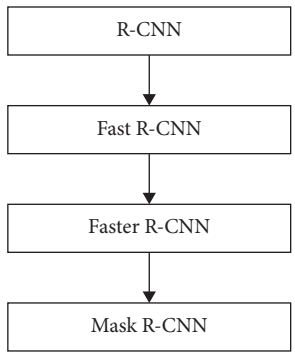

Figure 1: The development process of R-CNN algorithm, from R-CNN to Mask R-CNN.

detection speed but low detection precision and poor anti-interference ability. From the detection speed of CNN, Faster R-CNN is slower than YOLO and other one-stage algorithms when using GPU, but it can also reach the detection speed of 5FPS [31], which can fully meet the security requirements of X-ray machines.

2.1. Mask R-CNN. Based on Faster R-CNN, Mask R-CNN introduced the segmentation branch, which is composed of four convolutions, one deconvolution, and one convolution to realize instance segmentation. Moreover, ROI Align is proposed to fix the misalignment problem of ROI pooling, which could greatly improve the segmentation accuracy. The backbone of Mask R-CNN is ResNet [32] and Feature Pyramid Networks (FPN) [33]. The backbone uses residual learning to precisely extract object features and uses the feature pyramid to fuse multiscale features so as to construct high-quality feature maps.

After feature map extraction, RPNs are applied to extract ROIs from the feature maps. Then the ROIs are aligned and pooled by ROI Align. The aligned ROIs are used to instance segmentation by convolution and fully connected networks. The structure of MaskR-CNN is shown in Figure 2.

Mask R-CNN has many applications in image segmentation [34]. Mask R-CNN incorporates the advantages of previous algorithms and improves them to make the recognition more accurate, the training speed faster, and the effect better. In particular, the feature extraction structure of ResNet residual network + FPN is introduced, which solves the problem of difficult detection of small objects and has been applied in many fields. Mask R-CNN has significantly improved the effect on small object detection [35].

2.1.1. Feature Extraction Network. Mask R-CNN uses a feature extraction network composed with a feature pyramid structure by the residual network [36], which is divided into top-down and bottom-up parts. ResNet network is used in the bottom-up path, and five feature maps with different coarse granularities are generated through C_1 to C_5 modules. Each module is composed of multiple residual learning structures. For example, the structure of C_3 is $\{1\times1,128;\ 3\times3,128;\ 1\times1,512\}\times4$. This means that four residual learning structures are included, each of which consists of 3 convolution layers, with convolution kernels of 1×1, 3×3 and 1×1, and the number of channels of 128, 128, and 512, respectively. In the bottom-up path, the step size of the first convolution kernel in each stage is 2, and the step size of other convolution kernels is 1. Therefore, the size of the feature map is halved every time it passes through a module, and feature maps with different sizes can be obtained in this way.

In the top-down path, the high-level features are sampled twice, then the features obtained by 1×1 convolution are fused with those obtained by the bottom-up path, and finally a new feature map P2~P5 is obtained by 3×3 convolution. As shown in Figure 3, when an image with a size of 1024×1024 is input, the final feature map size is $\{32\times32,\ 64\times64,\ 128\times128,\ 256\times256\}$.

2.1.2. Region Proposal Network. Region Proposal Network [37] (RPN) is a full convolution neural network that uses the feature map to calculate the position of objects in images and can accept images of different sizes as input.

Different from the traditional Selective Search [38], the input of the RPN network is the feature map obtained by the feature extraction network. As shown in Figure 4, a multiscale anchor is generated by a sliding window in the feature map.

Further, as shown in Figure 5, RPN regresses each feature vector in the feature map to obtain a correction vector to correct the anchor. The correction value includes two confidences of foreground and background and four-position information, among which the correction mode of position information is shown in formula (1). Windows are generally represented by four-dimensional vectors (x, y, w, h). x, y is the coordinate of the center point, and w and h are the width and height of the candidate bounding box. By moving and zooming, the candidate bounding box is closer to its real position.

A large number of candidate bounding boxes can be obtained after the correction of anchor, and the foreground and background scores of these candidate bounding boxes are calculated, and the more accurate candidate bounding boxes are filtered out by nonmaximum suppression (NMS) [39].

$$\begin{cases} x = (1+\Delta x)\cdot x, \\ y = (1+\Delta y)\cdot y, \\ w = \exp(\Delta w)\cdot w, \\ h = \exp(\Delta h)\cdot h. \end{cases} \tag{1}$$

2.1.3. ROI Align. Since there is a certain correspondence between the image to be detected and the feature map of the image, mapping the target region in the image to the feature map is called a region of interest (ROI) mapping. In the Faster R-CNN algorithm, this process is completed in the ROI pooling layer, which converts input images of different sizes into a fixed dimensional feature vector output for

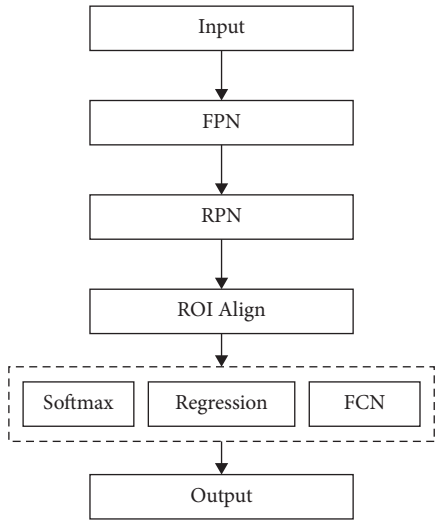

Figure 2: Mask R-CNN algorithm flow.

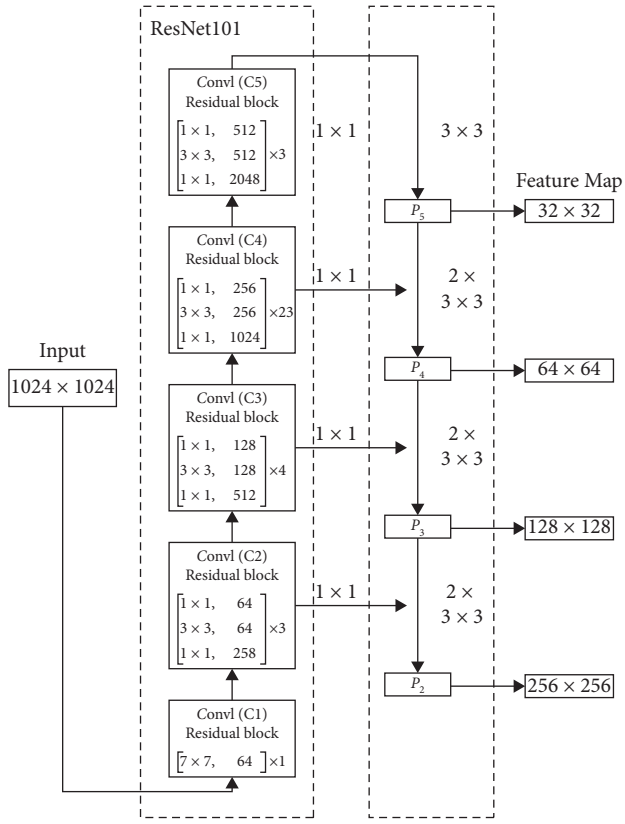

Figure 3: Feature extraction network using ResNet101.

subsequent head network training. The ROI Pooling uses the nearest-neighbor interpolation for region scaling, so if floating-point numbers are encountered in the process of pooling, they will be rounded off, resulting in the loss of information and then affecting the accuracy of detection.

The Mask R-CNN algorithms use ROI Align instead of ROI Pooling to avoid the loss of information. As shown in Figure 6, ROI Align uses bilinear interpolation in the process of region scaling, and the specific steps are traversing each candidate region, retaining the region boundary of floating-point numbers and dividing the candidate region into $k \times k$ cells, and the boundary of each cell also retains floating-point numbers.

The bilinear interpolation is used to calculate the fixed four candidate regions and divide them into $k \times k$ cells, and the boundaries of each cell are also preserved as floating-

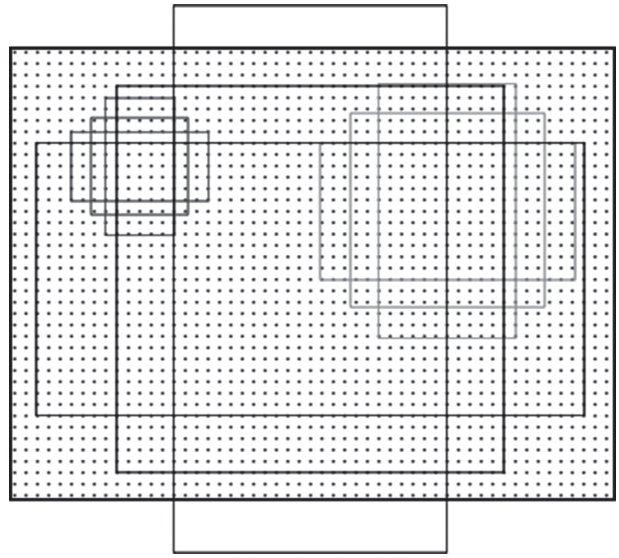

Figure 4: A multiscale anchor, generated by sliding window.

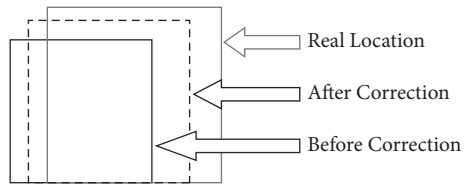

Figure 5: Correction process, the candidate bounding box is closer to its real position.

point numbers. Finally, the bilinear interpolation is used to calculate the values of fixed four coordinate positions, and then the max-pooling is carried out. ROI Align solves the problem of information loss in ROI Pooling by introducing bilinear interpolation for pooling, which turns the original discrete pooling process into a continuous process.

2.1.4. Network Header. Similar to Faster R-CNN, the network header of Mask R-CNN calculates the input characteristics of ROI Align. A fully connected layer plus Softmax is usually used to classify region proposals. At the same time, the region proposal is modified by two-stage regression in the same way as in the RPN process.

In addition, on the basis of classification and regression, Mask R-CNN obtains the accurate position information of the object through an FCN segmentation network [40] according to the obtained object border.

As shown in Figure 7, the input of the FCN segmentation network is the 14×14 characteristic map output by ROI Align, and the original 14×14 size is maintained using four 3×3 convolutional layers, then the size is boosted to 28×28 by a 2×2 deconvolution layer, and finally the 28×28 output is obtained by a 1×1 convolutional layer with a sigmoid activation layer. Each point in the output is the confidence of foreground and background in the region proposal, and the points are classified with a threshold of 0.5 to finally get the precise region of the target object.

2.2. Optimization of Mask R-CNN Algorithm. The process of using the Mask R-CNN algorithm to detect threat objects from an X-ray image is shown in Figure 8. Through network optimization, IoU index optimization, optimizer improvement, and prediction optimization of Mask R-CNN, the accuracy of threat object detection through X-ray by Mask R-CNN is improved.

2.2.1. Network Optimization. (1). Feature Extraction Optimization. In the feature extraction network, the high-level features focus on the overall object and the low-level features focus on the texture of the object, and the object can be better detected and localized by the low-level features. In the original feature extraction network shown in Figure 3, much information is lost in the lower layers, and the local texture of the object plays a greater role in the effectiveness of object detection due to the particularity of the X-ray image. Therefore, the original feature extraction network has room for improvement. Referring to the feature extraction network in PANet proposed in [41], as shown in Figure 9, this paper adds a bottom-up enhancement path after the top-down feature network in Figure 3 and fuses the low-level features into the high-level features again to avoid the loss of information.

(2) Online Hard Negative Example Mining. Traditional RPN network needs to take out ROI according to positive and negative samples of 1 : 3, in which the judgment condition of

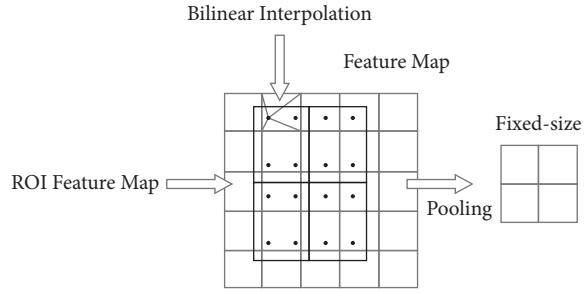

Figure 6: ROI Align, using bilinear interpolation to avoid the loss of information.

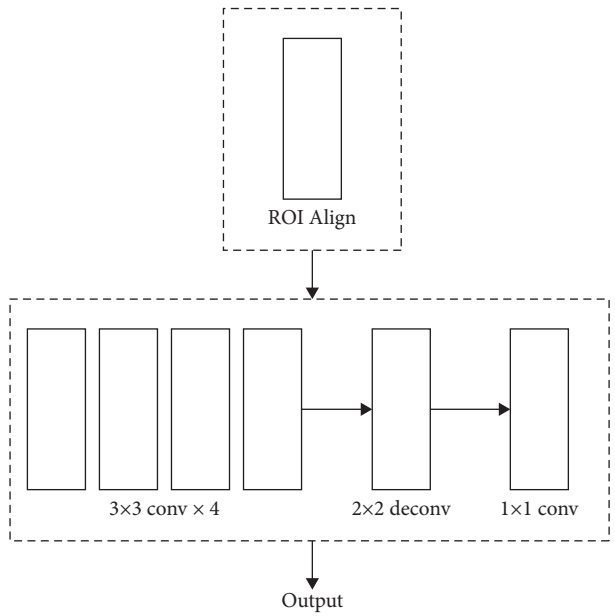

Figure 7: FCN segmentation network, output by ROI Align.

positive samples is that IoU is greater than 0.5, and the judgment condition of negative samples is that IoU is between [0.1, 0.5). The reason why it needs to be greater than 0.1 is to do a heuristic mining of hard examples. However, compared with online hard example mining (OHEM) [42], this heuristic hard example mining only uses prior hyperparameters, and OHEM can mine online hard examples, so it is more suitable for the iterative training mode of Mask R-CNN.

In the Mask R-CNN algorithm, ROI Align and its subsequent networks are called ROI networks. OHEM accomplishes online hard example mining by building two ROI networks. The idea is as follows: in the model training network, more ROIs are generated by the RPN network.

As shown in Figure 10, these ROIs are first passed to an OHEM network the same as the original ROI network, and the information of this network is only passed forward. The OHEM network calculates the loss of all ROLs and then sorts the ROIs by Loss, selects the specified number of ROIs with larger loss, and passes these ROIs to the ROI network for model training. Obviously, compared with selecting positive and negative samples through prior parameters, this online hard example mining can filter out more hard examples, and the robustness of the model can be improved by this training method.

(3) Edge Detection. X-ray image labeling often cannot be accurately labeled to the edges of objects, and most cases are labeled with rectangular boxes. Therefore, when using the Mask R-CNN algorithm directly, the accuracy of the Mask part will be affected. That is, the calculated L_{Mask} cannot accurately reflect the actual situation. In order to reduce the error, the Sobel operator is introduced before calculation, and the $Mask_{new}$ and GT_{new} are generated by the Sobel operator interacting with Mask and GT (Ground Truth) after Gaussian filter denoising.

Sobel operator is a commonly used edge detection operator [43], which interacts with Mask by using the Sobel operator. Sobel operator contains two sets of 3×3 matrices, representing horizontal and vertical, respectively. The horizontal and vertical luminance difference approximations can be obtained by plane convolution with the image. *Gx* and *Gy* represent the image gray values of edge detection by horizontal and vertical, respectively.

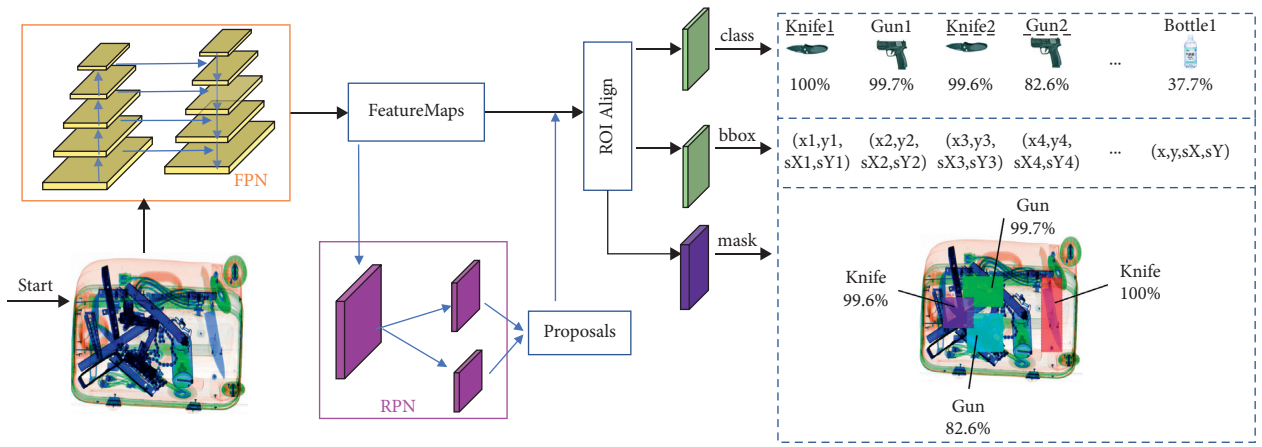

Figure 8: Mask R-CNN structure, from the input image to object classification and prediction.

The derivatives are derived in both directions as shown in equations (2) and (3), and then for each point of the image, the approximate gradient is derived as shown in equation (4).

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * \text{Mask}, \tag{2}$$

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * \text{Mask}, \tag{3}$$

$$\text{Mask}_{\text{new}} = G = \sqrt{G_x^2 + G_y^2}. \tag{4}$$

Sobel operator can make the features of edges in Mask and GT more prominent while weakening the nonedge part, so using the processed Mask_{new} and GT_{new} to calculate L_{mask} can make the Mask part fit more accurately.

2.2.2. IoU Index Optimization. In the traditional Mask R-CNN algorithm, IoU is used as a measure of the overlap degree of candidate regions. By calculating IoU, positive and negative samples can be determined, and the distance between candidate regions and the target objects can also be evaluated. Its advantage is that it is insensitive to scale and is not affected by the image to be detected and the size of the object.

However, IoU still has some limitations. For example, IoU cannot accurately reflect the coincidence degree between the two regions. As shown in Figure 11, when two regions overlap in different ways, the same IoU may be obtained. Obviously, the two regions in the first group have high overlap and better regression, while the last group has low overlap and poor regression.

To avoid these problems, DIoU (Distance-IoU) [44] can be used instead of IoU, which is more suitable for regression of candidate regions than traditional IoU. DIoU takes into account the distance, overlap rate, and scale between the object and the candidate region, and the regression of the candidate region can be more stable by using DIoU. The definition of DIoU is shown in the following formula:

$$\text{DIoU} = \text{IoU} - \frac{\rho^2\left(b, b^{gt}\right)}{c^2}, \tag{5}$$

where $\rho\,(b, b^{gt})$ represents the Euclidean distance between the center points of the two regions and c represents the diagonal distance of the minimum closed region containing both candidate regions, as shown in Figure 12. Compared with the traditional IoU, DIoU can directly minimize the distance between the center point of the prediction box and the real box, accelerate convergence, reduce errors, make the regression effect better, and make the results obtained in the process of SoftNMS more reasonable and effective.

2.2.3. Optimizer Improvement. When using the existing algorithms, the parameters of the model are updated by Mini-Batching. Because the image takes up too much video memory, in most cases, the hardware conditions cannot meet the requirement of training all samples at the same time. The idea of Mini-Batching is that during the training process of the model, each Epoch disrupts all the samples and generates multiple fixed-size subsets (Mini-Batch). The optimizer trains one Mini-Batch at a time and averages and updates the parameters trained by each sample in this Mini-Batch to the model. When the size of Mini-Batch is too small, this method almost degenerates to stochastic gradient descent (SGD) [45]. Compared with object detection in other scenarios, X-ray images tend to have fewer positive examples, which can lead to very few positive examples trained in each Mini-Batch when the size of the Mini-Batch is too small, resulting in greater randomness in model optimization.

To reduce the effect, this paper introduces a hyperparameter (integer n). In the training process of the model, the weight updating of the model is not carried out after each training of Mini-Batch. When n Mini-Batches are trained, the mean value of the weights in multiple Mini-Batches is calculated, and the model

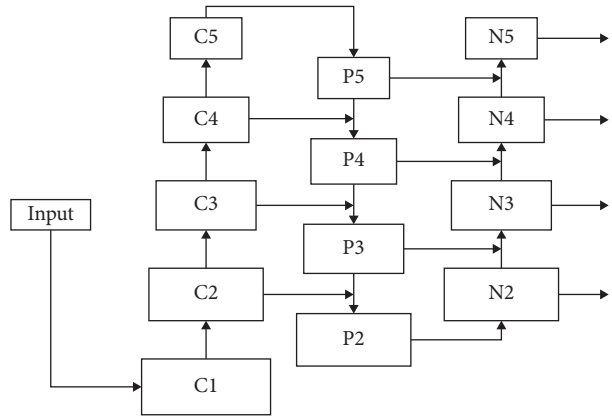

Figure 9: Improved feature extraction network from N2 to N5.

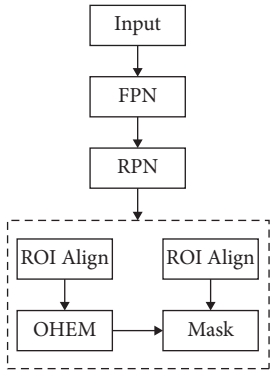

Figure 10: Training network with OHEM.

weights are updated. The purpose of expanding the size of Mini-Batch is achieved by this method.

2.2.4. Prediction Optimization. In the process of RPN, after the anchor is modified, the candidate regions need to be filtered by the NMS algorithm, and all the candidate regions are sorted according to their scores. The candidate region A_0 with the highest score is selected, and the intersection of union (IoU) of A_0 with another candidate region A_i is calculated. For A_i, whose IoU is higher than the specified threshold, it will be deleted, and then the one with the highest score is selected from the remaining candidate regions to repeat this step until no candidate regions can be deleted.

By this method, redundant candidate regions can be eliminated and the best detection position can be found. However, the NMS algorithm still has some limitations. On the one hand, it is a challenge to set a suitable threshold value. If the threshold value is set too high, fewer candidate regions will be removed, resulting in the ineffectiveness of the NMS algorithm; on the contrary, if the setting is too low, more candidate regions may be deleted, resulting in missed detection. On the other hand, if there are two objects with high coincidence in the image to be detected, the NMS algorithm will delete the objects with relatively low scores, so in this case, it is very likely to cause the missed detection of threat objects. The reason for this problem is that the NMS algorithm only considers the coincidence degree of other candidate regions and the candidate regions with the highest score but does not pay attention to the scores of these candidate regions. Through the recognition precision of NMS and SoftNMS [46], SoftNMS improves the effect significantly. To solve this problem, SoftNMS is used instead of the NMS algorithm, as shown in formula (6), where S_i is the classification confidence, N_t is the threshold, and M is the box with the highest confidence.

$$S_i = \begin{cases} S_i, & \mathrm{IoU}(M, b_i) < N_t, \\ 0, & \mathrm{IoU}(M, b_i) \geq N_t. \end{cases} \tag{6}$$

In the NMS algorithm, the candidate region that overlaps with the highest scoring candidate region M will be deleted with a score of 0. In the SoftNMS algorithm, the score of the candidate region is deducted by multiplying the score of the candidate region by Gaussian weight, as shown in the following formula:

$$S_i = S_i e^{-iou(M, b_i)^2/\sigma}. \tag{7}$$

When the overlapping candidate region has a high score, this method can avoid the candidate region being filtered out.

Taking the knife as an example, when the image to be detected is shown in Figure 13, the part marked by the color box is the threat object, and the two knives marked by the green box and blue box on the far left have more overlap. The result of the detection using the NMS algorithm is shown in Figure 14(a), and the knife with a lower score on the left side (corresponding to the green box in Figure 13) is filtered by the NMS algorithm, resulting in missed detection.

The detection results after replacing the NMS algorithm in Figure 14(a) with the SoftNMS algorithm are shown in Figure 14(b). Although the prediction probability of the missed detection of knives caused by the NMS algorithm is lower, they are not filtered out by the SoftNMS algorithm. Therefore, in this way, under the condition of ensuring the same detection precision of other items, the problem of missed detection by the NMS algorithm when there are multiple overlapping objects in the image can be effectively avoided.

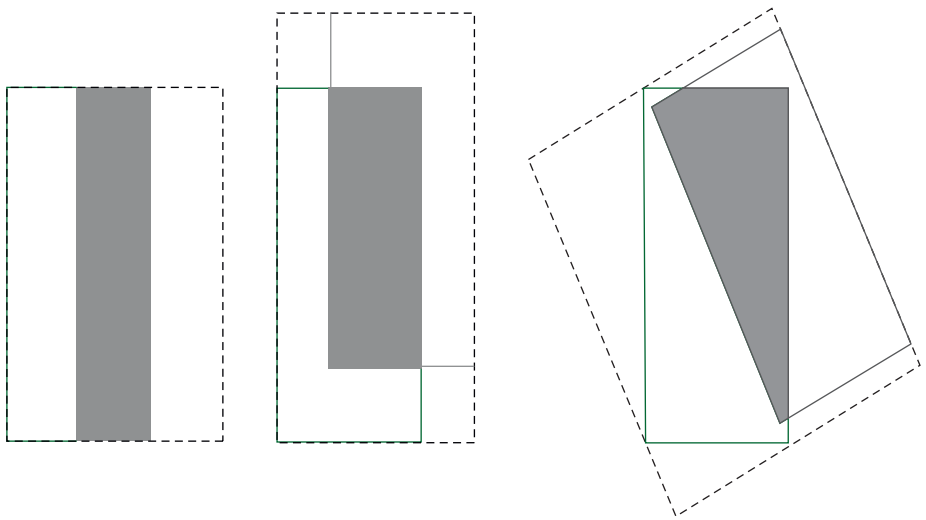

Figure 11: Different overlapping ways of the same IoU.

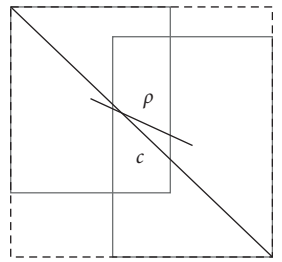

Figure 12: DIoU loss for bounding box regression.

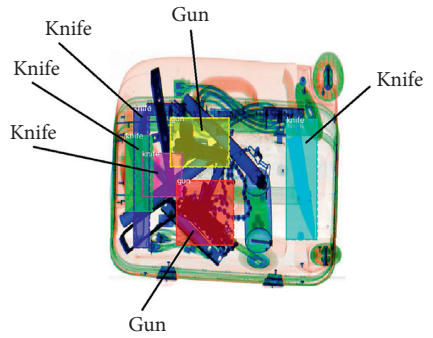

Figure 13: An example with ground truth.

3. Results and Discussion

3.1. Preprocessing of X-Ray Image Data. The original X-ray image contains many redundant parts, such as the text in the upper and lower parts and the blank background in Figure 15(a), but only the region marked by the dashed line in the figure is useful for algorithm training and prediction.

If the original image is directly used for algorithm training, on the one hand, there will be redundant noise which may lead to an inefficient training effect, and on the other hand, the size of the image passed into the training will be too large to slow down the speed of the algorithm training. Therefore, it is necessary to preprocess the original image before accessing the algorithm training and image analysis module. The contour detection algorithm in OPENCV can effectively identify the cargo in the original image, and the image processed by the algorithm is shown in Figure 15(b). Since the model training is implemented on the GPU, the redundant parts of the image are eliminated after preprocessing, the occupied video memory becomes smaller, and more images

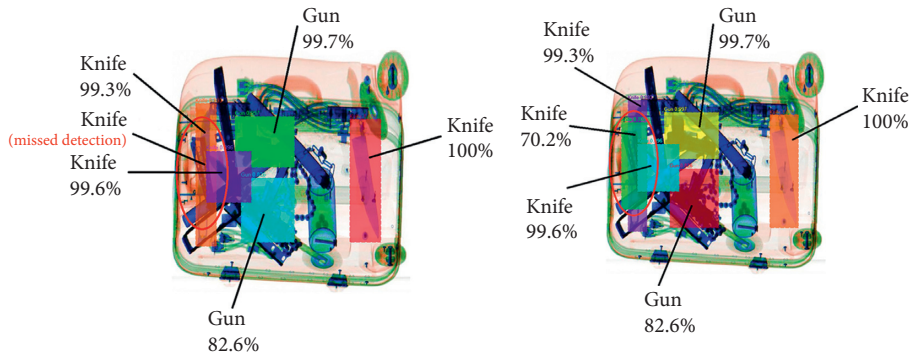

Figure 14: Detection results using NMS algorithm and SoftNMS algorithm. (a) Result of NMS. (b) Result of SoftNMS.

Figure 15: X-ray image preprocessing. (a) Original X-ray image. (b) Preprocessed image; the blank area was removed.

can be trained each time, so the effectiveness and efficiency of the algorithm training can be improved.

3.2. Experimental Environment and Parameter Settings. The hardware used in the experiment is shown in Table 1, and the software environment is shown in Table 2:

The X-ray image dataset is shown in Figure 16. We chose images from SIXray dataset [47] and simulated threat object images from the X-ray machine. It contains 5 classes of threat objects: knife, gun, liquid, mobile phone, and portable battery, with 6155 images in the train set and 560 images in the validation set. The algorithm parameters are set to train 2 images each time, and each Epoch contains 1000 trainings. The learning rate of the first 120 Epoch is 0.001, and then it is adjusted to 0.0001. The final algorithm achieves the best convergence effect at 160.

In order to verify the effectiveness of the improved algorithm, Mask R-CNN and other object detection algorithms are used to train the same dataset, and the training results are compared.

3.3. Comparison between Improved Algorithm and Mask R-CNN. In order to verify the effectiveness of the improved parts, the training results of the Mask R-CNN algorithm are compared with the detection results of the improved algorithm, and the results are shown in Figure 17.

The horizontal axis in Figure 17 is the recall rate, the vertical axis is the precision rate, and the three curves represent the detection effect under different IoU thresholds. The area below the curve is the AP value. It can be seen that in the results of the two algorithms, liquid, gun, and cell phone can be detected effectively. This is because the features of liquid, gun, and cell phone are obvious, and the objects and background are easy to distinguish, so recall and precision are both high; the knife and the portable battery have simple shapes, no obvious features, and it is not easy to distinguish between objects and background, so recall and precision are low. In the original Mask R-CNN algorithm, the detection effect of the knife and the portable battery are poor, while in the improved algorithm, the detection effect has been significantly improved.

It can be seen that the recall rate of the two classes of items, knife, and portable battery, is significantly increased

Table 1: Hardware environment.

CPU	Inter (R) Core (TM) i7-2600 CPU @3.40 GHz
GPU	TitanX 12G
Memory size	32 GB

Table 2: Software environment.

Operation System	CentOS 7
Development language	Python 3.6
Deep-learning API	Keras 2.0
Deep-learning framework	TensorFlow 1.3

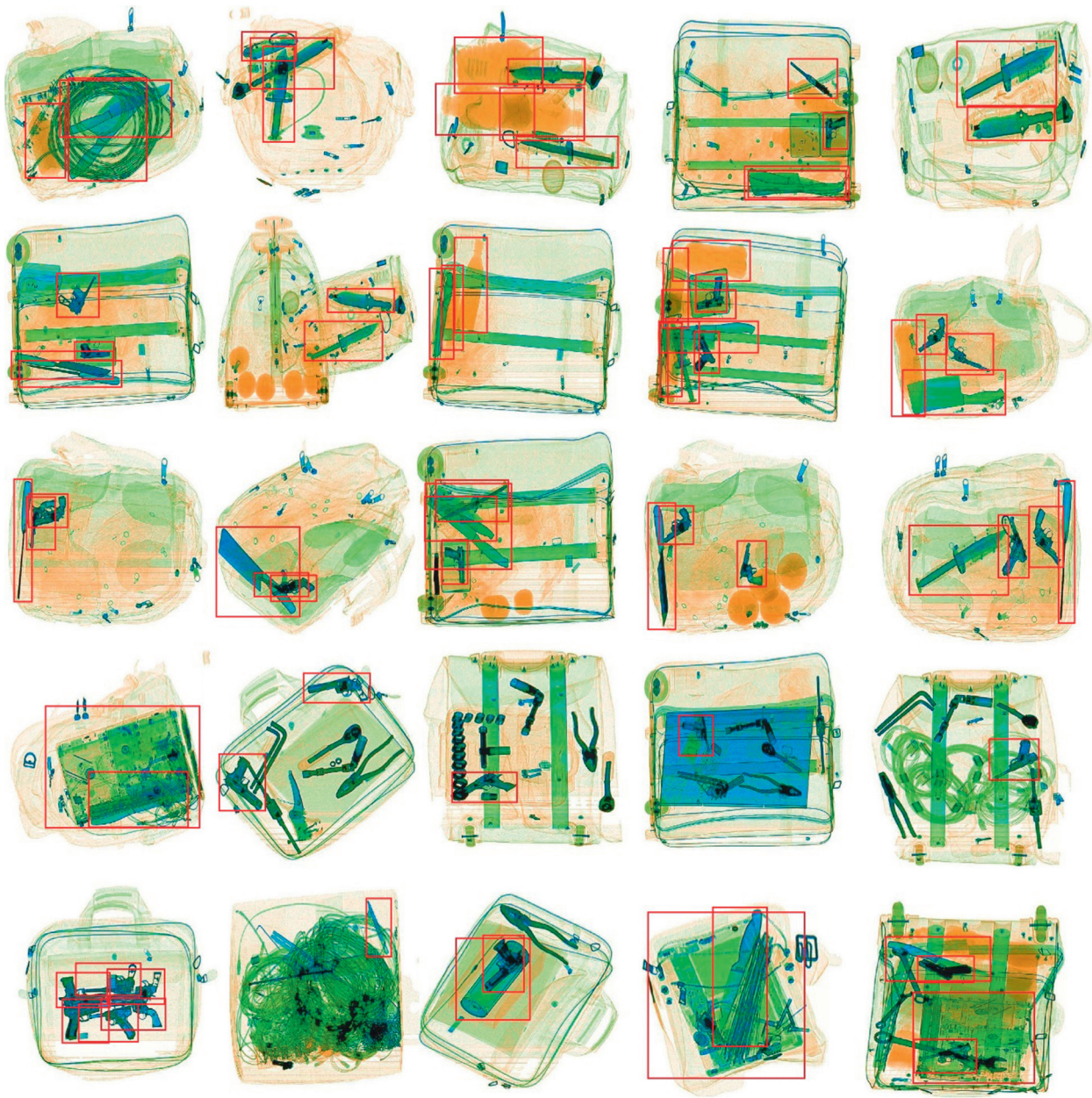

Figure 16: X-ray image dataset, including 5 classes of threat objects: knife, gun, liquid, mobile phone, and portable battery.

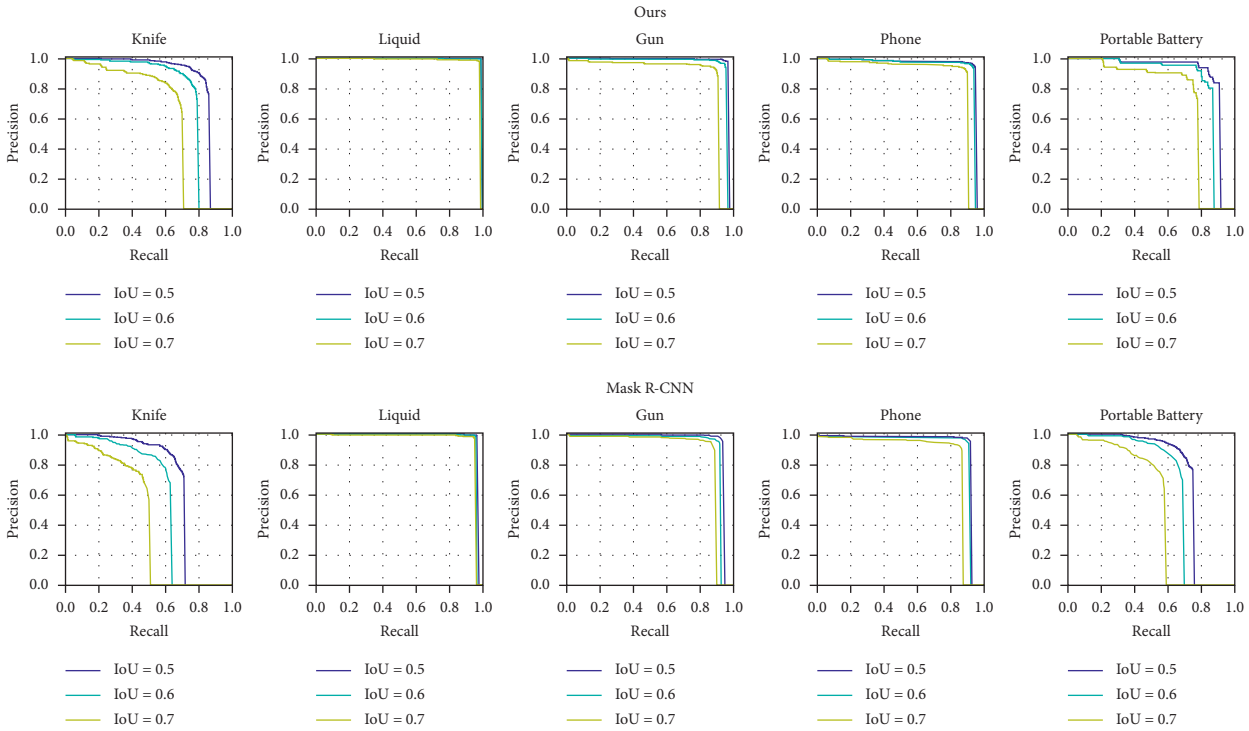

Figure 17: Comparison of detection effects of multiple objects.

in the improved algorithm (shown in the figure as the right shift of the intersection of the curve with the X-axis), and the accuracy rate is also obviously improved (shown in the figure as the upward shift of the curve, i.e., the increase of the value of Y for the same value of X).

From the aspect of algorithm structure, the improved algorithm has made a variety of optimizations in the network part (adding PANet enhancement path, introducing online hard example mining, introducing edge detection operator, optimizing IoU index, and improving optimizer), which has significantly improved the extraction effect of feature information and training effect of the algorithm. At the same time, SoftNMS is used to replace NMS in the inference part of the model, which reduces the cases where the detection objects are filtered out due to overlap and thus improves the recall rate of the algorithm model.

3.4. Comprehensive Comparison. The detection effect of the improved algorithm is compared with Mask R-CNN, Faster R-CNN, YOLOv3, and SSD513, and AP50 is used as the criterion. The results are shown in Table 3.

As can be seen, the detection effect of the one-stage algorithm (YOLOv3 and SSD513) is obviously lower than that of the two-stage algorithm due to its relatively simple network. The Mask R-CNN is improved based on Faster R-CNN, so the detection effect is better than that of Faster R-CNN, which is also a two-stage algorithm. Compared with the original Mask R-CNN algorithm, the improved algorithm in this paper has a significant improvement in detection effect because of the various optimizations mentioned above. The improved Mask R-CNN algorithm increases the value of mAP by 9.32%, and the AP values of the knife and portable battery with poor detection effect in the original Mask R-CNN algorithm is increased by 13.41% and 15.92%, respectively.

Table 3: Comparison of algorithm recognition results (AP50).

Threat objects						
Algorithm	Knife (%)	Liquid (%)	Gun (%)	Phone (%)	Portable battery (%)	Mean (%)
Ours	83.75	99.01	96.74	93.74	88.74	92.40
Mask R-CNN	70.34	94.43	90.22	87.60	72.82	83.08
Faster R-CNN	45.74	90.63	86.78	73.77	71.86	73.76
YOLOv3	39.16	90.12	85.77	70.03	69.53	70.92
SSD513	38.51	84.35	76.74	63.42	61.92	64.99

4. Conclusions

In order to realize the intelligent manufacturing of X-ray machine equipment, it is necessary to improve the accuracy of object detection. An improved algorithm based on Mask R-CNN is proposed in this paper, aiming at the problems of irregular placement, occlusion and overlap, small size, and simple shape in X-ray security inspection images.

(1) We optimized the network layer: bottom-up enhancement paths are added to fuse the features of the lower layers into the higher ones; we used OHEM to improve the robustness of the model. The training effect of sample model when accurate labeling is not

possible is improved by adding an edge detection module.

(2) We used DIoU instead of IoU to make the coincidence degree of object region and candidate region higher and the regression effect better. We selected the SoftNMS algorithm to replace the original NMS algorithm, which increased the object detection rate in the overlapping area of threat objects.

(3) We made an X-ray images dataset which included 5 classes of threat objects: knife, gun, liquid, mobile phone, and portable battery. The experimental results showed that the improved Mask R-CNN algorithm increases the mAP value by 9.32% compared with the original Mask R-CNN algorithm, and the AP values of knife and portable battery with poor detection effect increase by 13.41% and 15.92%, respectively.

(4) The proposed algorithm compared with other advanced algorithms such as Faster R-CNN, YOLOv3, SSD513. The results also indicated that the improved Mask R-CNN accomplished the most accurate precision attaining a mean accuracy precision of 92.40% with the test data set.

In summary, the results show the effectiveness and robustness of our proposed algorithm for threat object detection in X-ray images. Therefore, more research will be conducted to improve the accuracy of the small object such as knives and portable batteries with relatively low AP values.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this study.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant no. 51605069).

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[2] P. Chriskos, C. A. Frantzidis, P. T. Gkivogkli, P. D. Bamidis, and C. Kourtidou-Papadeli, "Automatic sleep staging employing convolutional neural networks and cortical connectivity images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 1, pp. 113–123, 2020.

[3] H. Kang, "Accelerator-aware pruning for convolutional neural networks," *In IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2093–2103, 2020.

[4] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 4, pp. 2848–2864, 2020.

[5] J. Deng, J. Guo, and S. Zafeiriou, "Single-stage joint face detection and alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1836–1839, Seoul, Korea, October 2019.

[6] M. Yousefi and J. H. L. Hansen, "Block-based high performance CNN architectures for frame-level overlapping speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 28–40, 2021.

[7] X. Du, T. Kurmann, P.-L. Chang et al., "Articulated multi-instrument 2-D pose estimation using fully convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 37, no. 5, pp. 1276–1287, 2018.

[8] S. Akçay, M. E. Kundegorski, M. Devereux, and T. P. Breckon, "Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 1057–1061, Phoenix, AZ, USA, September 2016.

[9] S. Akcay and T. P. Breckon, "An evaluation of region based object detection strategies within X-ray baggage security imagery," in *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1337–1341, Beijing, China, September 2017.

[10] S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, "Using deep convolutional neural network architectures for object classification and detection within X-ray baggage security imagery," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2203–2215, 2018.

[11] Y. Wei, Z. Zhu, H. Yu, and W. Zhang, "A real-time Threat Image Projection (TIP) model base on deep learning for X-ray baggage inspection," *Physics Letters A*, vol. 400, 2021.

[12] D. Mery, E. Svec, M. Arias, V. Riffo, J. M. Saavedra, and S. Banerjee, "Modern computer vision techniques for X-ray testing in baggage inspection," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 4, pp. 682–692, 2017.

[13] R. Gao, Z. Sun, J. Huyan et al., "Small foreign metal objects detection in X-ray images of clothing products using faster R-CNN and feature pyramid network," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[14] Y. F. A. Gaus, N. Bhowmik, S. Akçay, P. M. Guillén-Garcia, J. W. Barker, and T. P. Breckon, "Evaluation of a dual convolutional neural network architecture for object-wise anomaly detection in cluttered X-ray security imagery," in *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Budapest, Hungary, July 2019.

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

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 21–37, Amsterdam, Netherlands, October 2016.

[17] L. Barba-Guamán, J. E. Naranjo, A. Ortiz, and J. G. P. Gonzalez, "Object detection in rural roads through

SSD and YOLO framework," *Advances in Intelligent Systems and Computing*, vol. 1, pp. 176–185, 2021.

[18] S. Narejo, B. Pandey, D. E. vargas, M. Ciro Rodriguez, and R. Anjum, "Weapon detection using YOLO V3 for smart surveillance system," *Mathematical Problems in Engineering*, vol. 2021, Article ID 9975700, 9 pages, 2021.

[19] L. Pang, H. Liu, Y. Chen, and J. Miao, "Real-time concealed object detection from passive millimeter wave images based on the YOLOv3 algorithm," *Sensors*, vol. 20, no. 6, p. 1678, 2020.

[20] A. Warsi, M. Abdullah, M. N. Husen, M. Yahya, S. Khan, and N. Jawaid, "Gun detection system using YOLOv3," in *Proceedings of the 2019 IEEE International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)*, pp. 1–4, IEEE, Kuala Lumpur, Malaysia, August 2019.

[21] H. Huang, D. Sun, R. Wang, C. Zhu, and B. Liu, "Ship target detection based on improved YOLO network," *Mathematical Problems in Engineering*, vol. 2020, Article ID 6402149, 10 pages, 2020.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014.

[23] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1440–1448, Honolulu, HI, USA, July 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] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, Santiago, Chile, December 2017.

[26] H. Nguyen, "Improving faster R-CNN framework for fast vehicle detection," *Mathematical Problems in Engineering*, vol. 2019, Article ID 3808064, 11 pages, 2019.

[27] X. Dai, J. Hu, H. Zhang et al., "Multi-task faster R-CNN for nighttime pedestrian detection and distance estimation," *Infrared Physics & Technology*, vol. 115, no. 4, 2021.

[28] Y. Liu, "An improved faster R-CNN for object detection," in *Proceedings of the 2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, pp. 119–123, Hangzhou, China, December 2018.

[29] X. Han, "Modified cascade RCNN based on contextual information for vehicle detection," *Sensing and Imaging*, vol. 22, no. 1, pp. 1–19, 2021.

[30] B. Liu, J. Luo, and H. Huang, "Toward automatic quantification of knee osteoarthritis severity using improved Faster R-CNN," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 3, pp. 457–466, 2020.

[31] F. Y. Hu, L. Y. Li, and X. R. Shang, "A review of object detection algorithms based on convolutional neural networks," *Journal of SuZhou University of Science and Technology (Natural Science Edition)*, vol. 37, no. 2, pp. 1–10, 2020.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Salt Lake City, UT, USA, June 2016.

[33] T. Y. Lin, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, Honolulu, HI, USA, July 2017.

[34] X. Fu, J. Wang, Z. Hu, Y. Guo, and R. Wang, "Automated segmentation for whole human eye OCT image using RM multistage Mask R-CNN," *Applied Optics*, vol. 60, no. 9, pp. 2518–2529, 2021.

[35] Y. Liu, "A survey of research and application of small object detection based on deep learning," *Acta Electonica Sinica*, vol. 48, no. 3, p. 590, 2020.

[36] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1492–1500, Boston, MA, USA, June 2017.

[37] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, 2017.

[38] Z. Fang, Z. Cao, Y. Xiao, K. Gong, and J. Yuan, "MAT: Multianchor visual tracking with selective Search region," *IEEE Transactions on Cybernetics*, vol. 2020, Article ID 3039341, 15 pages, 2020.

[39] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, pp. 850–855, 2006.

[40] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, Salt Lake City, UT, USA, June 2015.

[41] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8759–8768, Salt Lake City, UT, USA, June 2018.

[42] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 761–769, Las Vegas, NV, USA, June 2016.

[43] R.-G. Zhou and D.-Q. Liu, "Quantum image edge extraction based on improved Sobel operator," *International Journal of Theoretical Physics*, vol. 58, no. 9, pp. 2969–2985, 2019.

[44] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: faster and better learning for bounding box regression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12993–13000, 2020.

[45] L. Bottou, "Stochastic gradient descent tricks," *Lecture Notes in Computer Science*, vol. 7700, pp. 421–436, 2012.

[46] N. Bodla, B. Singh, R. Chellappa, and S. Larry, "Davis. "Soft-NMS -- improving object detection with one line of code," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5561–5569, Venice, Italy, October 2017.

[47] C. Miao, L. Xie, F. Wan et al., "SIXray: a large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2119–2128, Long Beach, CA, USA, June 2019.