Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/75/20240552

Contraband detection method in X-ray image based on improved CenterNet network

Dong Yang[1,2], Wen Tang[2,3], Chao Li[2], Xinghua Wu[2], Jianchao Wang[2]
[1]Graduate School of China Academy of Railway Science, Beijing 100081, China
[2]Institute of Computing Technology, China Academy of Railway Sciences, Beijing 100081, China

[3]1036139292@qq.com

Abstract. Security inspection has been played a crucial role in ensuring the safety of railway passenger transportation. The detection of contraband in X-ray images is an important part of safety inspection work. Since the X-ray image of security inspection is a pseudo color image with a unique imaging style, and the restricted objects are mostly small targets, the accuracy of the object detection algorithm using visible light images directly is not high. This research will propose an improved CenterNet method, and design a 16 times down sampling rate backbone network, which is more suitable for detecting contraband in X-ray images. This article will introduce attention mechanisms to increase attention to effective and key areas in X-ray images, and reveal the mechanisms of spatial and channel attention in X-ray images. Furthermore, by designing an adaptive feature fusion mechanism, the drawback of the original CenterNet by using only the last layer of output features is compensated. Experiments have shown that on the SIXray dataset, mAP is improved by 3.3% compared to the original CenterNet method, and 5.5%, 2.4%, 3.6%, 10.1% compared to SSD, Yolo V3, FPN, DERT algorithms. The detection accuracy of restricted items in X-ray images is effectively improved.

Keywords: X-ray Image, Railway Security Inspection, Detection of Contraband, Feature Fusion, CenterNet.

1. Introduction

Railway is one of the most important modes of transportation in China. It will bring threats to transportation security once a terrorist attack or violent incident occurs in railway stations or on the train, and, it may also cause great psychological trauma to the passengers. Therefore, government attaches great importance to railway safety and security inspection.

X-ray security detectors are widely used in railway passenger security inspection, providing critical support for ensuring passenger transportation safety [1],[2]. By utilizing the different attenuation characteristics of X-rays penetrating different objects, X-ray security detectors can generate X-ray images by scanning and imaging the items entering the detector in real time[3]. The X-ray image contains information about the shape and material of the scanned item. Therefore, by analyzing the X-ray image, it is possible to determine whether the corresponding package contains contraband and complete the security check.

At present, the discrimination of contraband in X-ray images in railway passenger security inspections mainly relies on manual work, and the quality of image recognition is easily affected by

DOI: 10.54254/2755-2721/75/20240552

factors such as the working condition of the security inspector. Therefore, it is critical to use deep learning, convolutional neural networks, and other technical means to assist in image recognition, which could reduce the labor intensity of image recognition personnel, and improve the quality of security inspection operations.

The detection of contraband in X-ray images refers to the detection of contraband in X-ray images through computer vision and object detection technology. Reference [4] disclosed the X-ray image dataset SIXray, which includes 8929 images with contraband labeled, on the basis of which it conducted research on the classification of contraband. Reference [5] conducted research on contraband detection based on the Cascade R-CNN method, and improved the method by introducing dynamic deformable convolution (DyDC) and adaptive IOU mechanism. Reference [6] conducted research on contraband detection based on Faster RCNN method, designed a candidate recommendation network, and introduced dilation convolution. Reference [7] conducted research on the detection of guns, a type of contraband, based on Faster RCNN, Mask RCNN, and RetinaNet algorithms. Reference [8] conducted research on contraband detection based on the YOLO-V3 network, and introduced the DioU loss function and spatial pyramid pooling feature method. Reference [9] conducted research on contraband detection based on the YOLO-V3 network, and improved the detection accuracy by introducing dense connections and improving the loss function. Reference [10],[11],[12] conducted research on contraband based on YOLO-V5 network, improving detection accuracy by using attention mechanisms, improving loss functions, and other methods.

Based on the above studies, the following problems exist with regard to the detection of contraband in the X-ray image: (1) Most researches consider the issue of object detection solely from the perspective of image processing, and lack analysis of the imaging mechanism of X-ray images and the characteristics of contraband; (2) The researches are mainly based on one-stage and two-stage target detection methods, and the algorithm accuracy and speed need to be further improved. This paper adopts an anchor-free network for the detection of restricted items. The main innovations are as follows: (1) Designing a 16 times down sampling rate backbone network, which is more suitable for the X-ray images, based on analysis of the imaging mechanism of X-ray images and the characteristics of restricted items; (2) Introducing attention mechanism to improve detection accuracy, and revealing the mechanisms of spatial attention mechanism and channel attention mechanism in X-ray images; (3) Introducing feature fusion methods to further improve detection accuracy.

2. Characteristics of the X-ray images

The X-ray image generated by the security detector is a pseudocolor image, usually consisting of three colors: blue, orange, and green. Blue represents inorganic matter, usually metal, orange represents organic matter, and green represents a mixture of the two. An example of X-ray image is shown in figure 1. The objects in blue are knives and guns, which are contraband to be detected, and the objects in green and orange are regular items. Compared to visible light images, the X-ray images have the following characteristics: (1) There is a white background in all four corners of the image, which does not contain any information; (2) Contraband such as knives and guns typically contain metal, meaning that the blue part of the image contains a greater amount of information; (3) Due to the fact that contraband are usually packed in suitcases and backpacks, the targets to be detected usually occupy a relatively small portion of the screen, and are mostly small items; (4) There is no logical correlation between the appearance and location of the various objects within the image. By contrast, such correlations do exist in visible light images: for example, humans are on the ground and birds are in the sky.

DOI: 10.54254/2755-2721/75/20240552

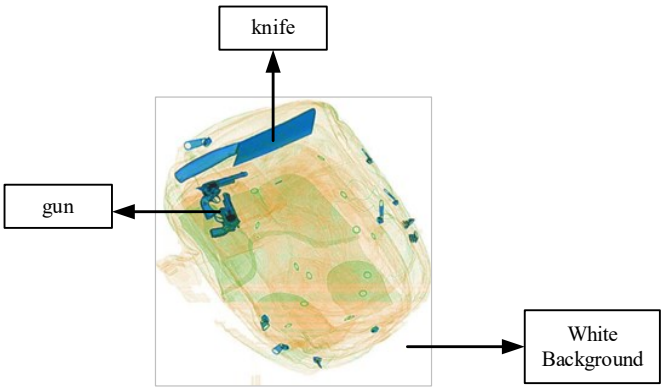

Figure 1. Example of X-ray image

3. Improved CenterNet Network

Zhou et al. proposed CenterNet in 2019. The core idea of the algorithm is to treat the target to be detected as a point. The algorithm labels the target by predicting the center point position, center point offset, and target width and height. There is no need to preset the anchor box in advance, which eliminates non maximum suppression operations. In this way, the algorithm can improve detection accuracy as well as detection speed. This article selects CenterNet as the baseline algorithm to conduct research on the detection of contraband in security X-ray images.

The improved CenterNet structure as shown in figure 2 is mainly composed of four parts: the improved ResNet18 network, CBAM residual block, adaptive feature fusion, and head network. (1) A residual network with 16 times down sampling rate was designed based on ResNet18 network, resulting in an improved ResNet18 network. (2) An attention mechanism is introduced. CBAM modules are added to the last convolutional layer of Convolutional Group 2, Convolutional Group 3, and Convolutional Group 4 in the improved ResNet18, forming CBAM residual blocks 2-2, CBAM residual blocks 3-2, and CBAM residual blocks 4-4. (3) An adaptive feature fusion method is designed to perform feature fusion on the inputs of CBAM residual block 2-2, CBAM residual block 3-2, and CBAM residual block 4-4. (4) Using the head network analyzes the fused features and outputs predictions.

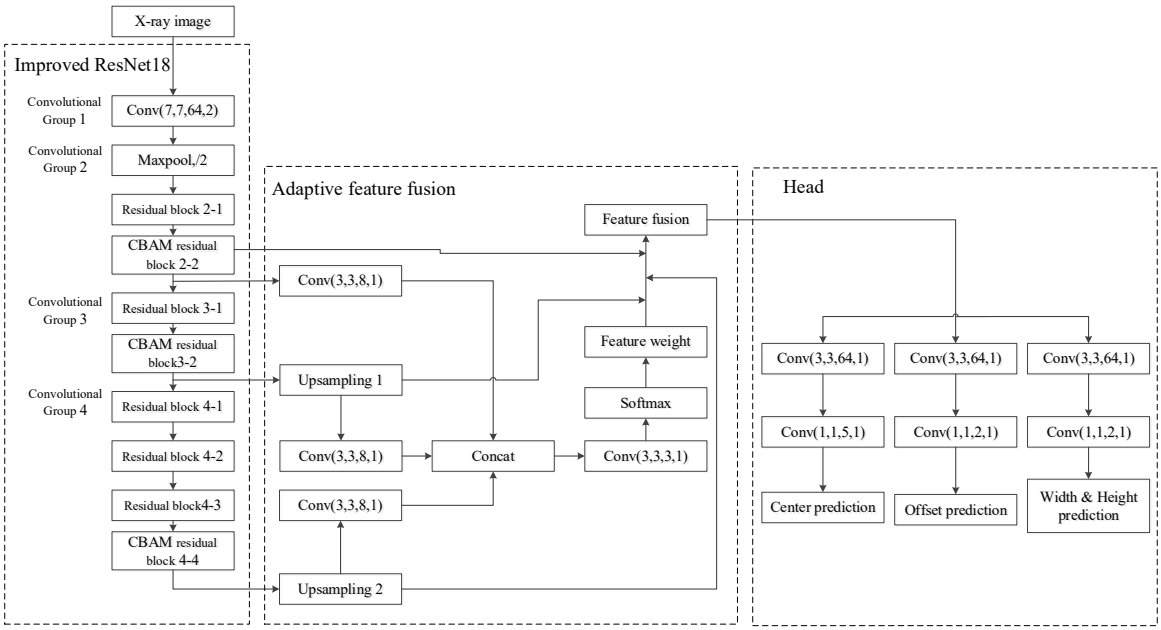

Figure 2. Improved CenterNet architecture

DOI: 10.54254/2755-2721/75/20240552

3.1. Improved ResNet18 backbone network

Regarding CenterNet, the author used the ResNet18 backbone network with a down sampling rate of 32 times. However, there is a significant difference between visible light images and security X-ray images. Through visualization research on convolutional layers in the dataset ImageNet, convolutional layers with different down sampling rates can extract different features. The convolutional layer with double down sampling can extract corner and edge features, the convolutional layer with triple down sampling can extract texture features, the convolutional layer with four times down sampling can extract features that are more closely related to specific categories, and the convolutional layer with five times down sampling can extract diverse features of the entire object[13]. In addition, from the perspective of down sampling on the perception field of feature maps, the perception field of high-level features is larger, which is helpful for detecting large targets, while that of low-level features is smaller, which is helpful for detecting small targets. Admittedly, removing the convolutional layer may cause discarding features of a certain scale, but too many convolutional layers may cause performance degradation at the same time. It follows that a backbone network with a down sampling rate of 32 times is suitable for object detection in visible light images, but not suitable for contraband detection in X-ray images.

This article adopts a backbone network with a down sampling rate of 16 times for contraband detection in security X-ray images. The improved ResNet18 network is shown in figure 3. The fifth convolutional group is removed from the ResNet18 network, with the result that the down sampling rate of the network changes to 16 times, and two residual modules are added after the fourth convolutional group in the ResNet18 network, which keeps the network still composed of 18 convolutional layers.

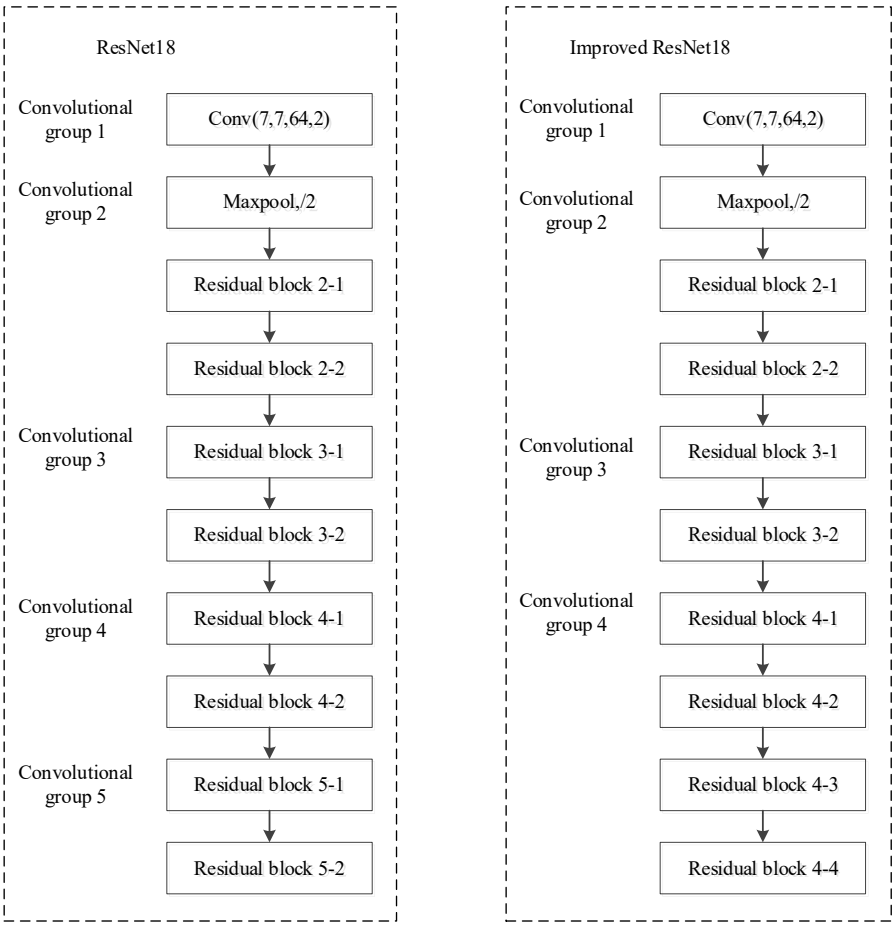

Figure 3. Improved ResNet18 network

DOI: 10.54254/2755-2721/75/20240552

3.2. Attention mechanism

Based on the imaging principle of X-ray images, in the X-ray image, the area without objects is imaged as white, and this area does not require attention; the area of metal material is imaged in blue, and, as most contraband such as knives and guns contain metal, the blue area needs to be focused on. In addition, the characteristic of X-ray images is that most of the items are overlapped. Therefore, attention mechanisms can be introduced to process feature layers, so as to enhance the focus of the network on key areas and improve detection accuracy. The CBAM module[14] is a lightweight module that comprehensively utilizes channel attention mechanism and spatial attention mechanism. By adjusting the weights of input features in channel and spatial dimensions, the attention mechanism is implemented. The module structure is shown in figure 4.

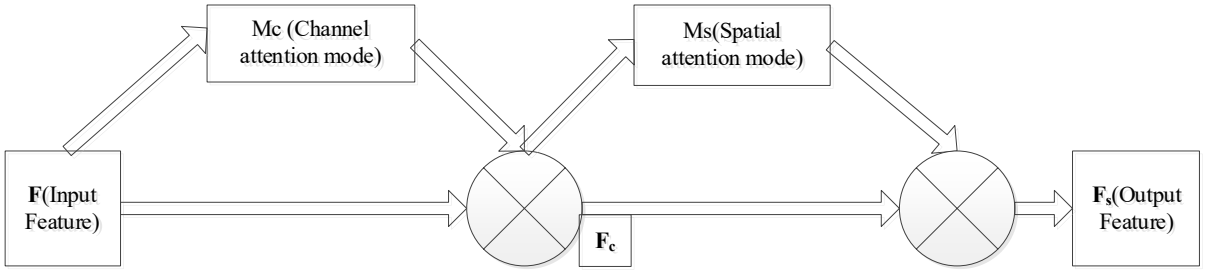

Figure 4. CBAM Structure

For input features$\boldsymbol{F} \in R^{C\times H\times W}$, the CBAM module first infers to generate a one-dimensional channel attention module $M_c \in R^{C\times 1\times 1}$, and then infers to generate a two-dimensional spatial attention module$M_s \in R^{1\times H\times W}$. For output feature $\boldsymbol{F}$, the entire process can be represented as:

$$\boldsymbol{F_c} = M_c(\boldsymbol{F}) \otimes \boldsymbol{F} \tag{1}$$

$$\boldsymbol{F_s} = M_s(\boldsymbol{F_c}) \otimes \boldsymbol{F_c} \tag{2}$$

The improved ResNet18 backbone network adds CBAM modules to the residual blocks 2-2, 3-2, and 4-4 of to form a CBAM residual block, enhancing the network's focus on effective and key areas in security images. The structure of the CBAM residual block is shown in figure 5. CBAM module is added to the two convolutional layers in the residual block. Feature x is processed by the two convolutional layers and the CBAM module to obtain F (x), which is then added to the feature x to obtain the final feature output.

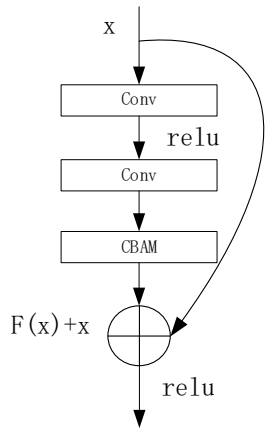

Figure 5. CBAM residual block structure

DOI: 10.54254/2755-2721/75/20240552

3.3. Adaptive feature fusion

CenterNet deconvolutes the output features of the last layer and restores them to a feature map that is 1/4 times the size of the input image. Using this feature map for target prediction will result in loss and waste of feature information contained in other feature layers, which is not conducive to the task of detecting contraband. The feature fusion method can fully utilize the features output from different convolutional layers and improve the quality of final output features. Various methods such as FPN[15], PANet[16], ASFF[17] have fully demonstrated the effectiveness of multi-scale features and feature fusion. Inspired by ASFF, the features output by CBAM residual blocks 2-2, 3-2, and 4-4 in the improved ResNet18 network are adaptively fused to enrich the information of the output features and improve detection accuracy.

The CBAM residual block 4-4 outputs a feature map that is 1/16 times the size of the input image, and uses two sets of 2-times-upsampling deconvolution to recover a feature map that is 1/4 times the size of the input image. The CBAM residual block 3-2 outputs a feature map that is 1/8 times the size of the input image, and uses a set of 2-times-upsampling deconvolution to recover a feature map that is 1/4 times the size of the input image; CBAM residual block 2-2 outputs a feature map that is 1/4 times the size of the input image. Then, the feature values of the same points in the three feature maps are weighted and summed point by point. The fusion method is as follows:

$$\boldsymbol{Y_{ij}} = \alpha_{ij}\boldsymbol{X_{ij}^{2}} + \beta_{ij}\boldsymbol{X_{ij}^{3to2}} + \gamma_{ij}\boldsymbol{X_{ij}^{4to2}} \tag{3}$$

where $\boldsymbol{Y_{ij}}$ is the vector of the fused feature map at position (i, j), $\boldsymbol{X_{ij}^{2}}$ is the vector of the feature map output from CBAM residual block 2-2 at position (i, j), α_{ij} is the weight coefficient of X_{ij}^{2}, $\boldsymbol{X_{ij}^{3to2}}$ is the vector of the feature map output from CBAM residual block 3-2 at position (i, j) after being restored to 1/4 of the input image size, β_{ij} is the weight coefficient of $\boldsymbol{X_{ij}^{3to2}}$, $\boldsymbol{X_{ij}^{4to2}}$ is the vector of the feature map output from CBAM residual block 4-4 at position (i, j) after being restored to 1/4 of the input image size, γ_{ij} is the weight coefficient of $\boldsymbol{X_{ij}^{4to2}}$. In order to ensure that the fused feature map remains normalized, it is required thatα_{ij}+β_{ij} + γ_{ij}=1. Therefore,α_{ij} 、β_{ij}、γ_{ij} are the values processed through the softmax function, defined as follows:

$$\alpha_{ij} = \frac{e^{\lambda_{\alpha ij}}}{e^{\lambda_{\alpha ij}} + e^{\lambda_{\beta ij}} + e^{\lambda_{\gamma ij}}} \tag{4}$$

$$\beta_{ij} = \frac{e^{\lambda_{\beta ij}}}{e^{\lambda_{\alpha ij}} + e^{\lambda_{\beta ij}} + e^{\lambda_{\gamma ij}}} \tag{5}$$

$$\gamma_{ij} = \frac{e^{\lambda_{\gamma ij}}}{e^{\lambda_{\alpha ij}} + e^{\lambda_{\beta ij}} + e^{\lambda_{\gamma ij}}} \tag{6}$$

where $\lambda_{\alpha ij}$、$\lambda_{\beta ij}$、$\lambda_{\gamma ij}$ are the original weight values learned by the convolutional network.

3.4. Loss function

The loss function consists of three parts: center point classification loss, center point offset loss, and target regression loss. The center point classification loss is similar to the Focal Loss[18] function, represented as follows:

$$L_k = -\frac{1}{N}\sum_{\text{xyc}} \begin{cases} \left(1-\boldsymbol{\hat{Y}_{xyc}}\right)^{\alpha} exp\left(\boldsymbol{\hat{Y}_{xyc}}\right), \boldsymbol{\hat{Y}_{xyc}} = 1 \\ \left(1-\boldsymbol{Y_{xyc}}\right)^{\beta}\left(\boldsymbol{\hat{Y}_{xyc}}\right)^{\alpha} exp\left(1-\boldsymbol{\hat{Y}_{xyc}}\right), others \end{cases} \tag{7}$$

where α and β are hyperparameters in the Focal Loss function, taken as 2 and 4 respectively. N is the total number of key points in the feature map. $\boldsymbol{\hat{Y}_{xyc}}$ is the predicted value, representing the probability

DOI: 10.54254/2755-2721/75/20240552

of the key points being the target center, and the value range is [0,1]. $\boldsymbol{Y_{xyc}}$ is the true value, using Gaussian kernel to smooth the probability near the center point, and the value range is [0, 1].

Because there is 4 times down sampling in the feature map, the predicted center point and the real value will be offset. The center offset loss adopts L1 loss function, which is expressed as follows:

$$L_{off} = \frac{1}{N}\sum_{p}\left|\widehat{\boldsymbol{O}}_{\widetilde{\boldsymbol{p}}} - \left(\frac{\boldsymbol{p}}{R} - \widetilde{\boldsymbol{p}}\right)\right| \tag{8}$$

where N is the total number of key points in the feature map, $\widehat{\boldsymbol{O}}_{\widetilde{\boldsymbol{p}}}$ represents the predicted center offset, $\boldsymbol{p}$ represents the coordinates of the center points in the original image, R is the down sampling rate, and $\widetilde{\boldsymbol{p}}$ represents the corresponding center points on the feature map.

Regression is performed on the width and height of the target, and the target regression loss L_{wh} adopts L1 loss function, which is expressed as follows:

$$L_{wh} = \frac{1}{N}\Sigma\left(\left|\boldsymbol{w_{pred}} - \boldsymbol{w_{gt}}\right| + \left|\boldsymbol{h_{pred}} - \boldsymbol{h_{gt}}\right|\right) \tag{9}$$

where $\boldsymbol{w_{pred}}$ is the width of the predicted target, $\boldsymbol{h_{pred}}$ is the height of the predicted target, $\boldsymbol{w_{gt}}$ is the width of the real target, $\boldsymbol{h_{gt}}$ is the height of the real target, and N is the total amount of data for width and height.

The final loss function is:

$$L = L_k + \lambda_1 L_{wh} + \lambda_2 L_{off} \tag{10}$$

where λ_1 is the target regression loss weight, taken as 0.1, and λ_2 is the center point offset loss weight, taken as 1.0.

4. Experiments and analysis

The experimental environment was Ubuntu 18.04, the GPU was RTX3080Ti, the algorithm implementation used the mmdetection framework [19], and the experimental dataset was selected from the public dataset SIXray.

4.1. Dataset and preprocessing

The SIXray dataset contains 8929 X-ray images with labeling information of contraband, including six categories: guns, knives, scissors, wrenches, pliers, and hammers. The number of labeled samples for each category is shown in table 1. Due to the scarcity of images containing hammer labeling (60 images), only five categories (guns, knives, wrenches, pliers, and scissors) were analyzed and tested. 70% of the images are randomly selected as the training set, totaling 6250 images. 30% of the images are randomly selected as the test set, totaling 2679 images.

Table 1. Sample Labeling Quantity of SIXray Dataset

	Gun	Knife	Wrench	Plier	Scissor	Hammer
Total	3131	1943	2199	3961	983	60

Due to the insufficient sample size in the SIXray dataset, the training images were randomly cropped, resized, normalized, and flipped sequentially to further enrich the samples.

4.2. Experimental parameters

The ResNet18 network uses a pre-trained model on the ImageNet dataset for transfer learning. To accelerate the training process, the improved ResNet18 network uses the first four convolutional groups of the pre-trained ResNet18 network on the ImageNet dataset for transfer learning.

DOI: 10.54254/2755-2721/75/20240552

The train Epochs are set to 90, the Batch Size is set to 16, and SGD is used as the optimizer with a learning rate of 0.003 and momentum of 0.9. The learning rates decay to 1/10 at the 60th and 80th Epochs respectively. The evaluation indicator is mAP defined in VOC2007.

4.3. Experimental Analysis

4.3.1. Comparative experiment on improved ResNet18 backbone network

To study the impact of different down sampling rates on the backbone network, a comparative experiment was conducted by selecting ResNet18, the first four convolutional groups of ResNet18, and the improved ResNet18 as the backbone network. CenterNet is chosen as the detection algorithm, and the output of the backbone network was restored to a feature map 1/4 times the input image size through deformable convolution and up sampling. The experimental results are shown in Table 2. The backbone network with a 16-fold down sampling rate is more suitable for detecting contraband in security X-ray images, and the detection accuracy of the improved ResNet18 is significantly enhanced.

4.3.2. Adaptive feature fusion

The X-ray image was input into the ResNet18 network and the output features of convolution group 2, convolution group 3, convolution group 4, and convolution group 5 were visualized. The results are shown in figure 6. The output features of convolution group 5 contain less useful information.

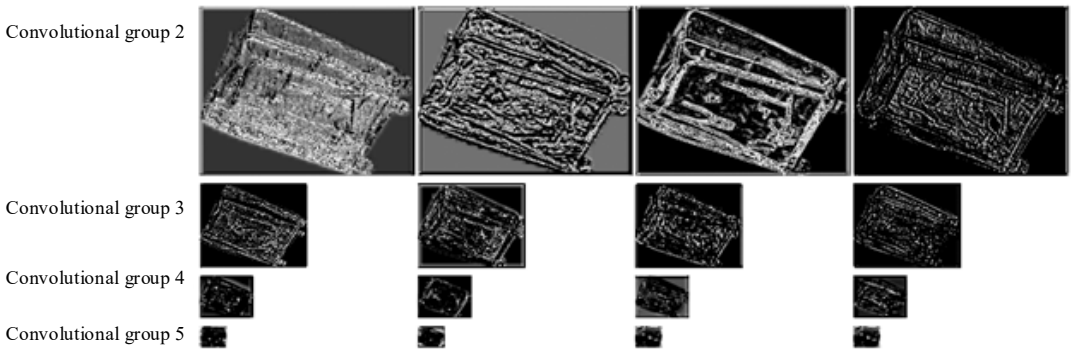

Figure 6. Output feature of each convolution group

In the improved ResNet18, there are four convolution groups. To study the detection accuracy of output features of different convolution groups after adaptive fusion, different combinations of output features of convolution groups 4, convolution groups 3, and convolution groups 2 were selected. Each convolution group was restored to a feature map of 1/4 times the input image size through deformable convolution and up sampling, on which adaptive fusion was performed. The experimental results are shown in table 2. The mAP after fusing the output features of convolution group 2 and convolution group 4 increased by 0.2% compared to convolution group 4 by itself, while the mAP after fusing the output features of convolution group 2, convolution group 3, and convolution group 4 increased by 0.8% compared to convolution group 4 by itself. It can be seen that the output of adaptive fusion of the three convolution groups is the optimal fusion method.

Table 2. Detection results of different feature fusion

Fusion mode	Convolutional group 2	Convolutional group 3	Convolutional group 4	mAP
Fusion mode 1			√	86.6
Fusion mode 2	√		√	86.8
Fusion mode 3	√	√	√	87.4

DOI: 10.54254/2755-2721/75/20240552

4.3.3. Comparative experiment of CBAM module

To investigate the impact of CBAM modules, a control experiment was conducted. Examples of common security X-ray images are shown in figure 7(a) and figure 7(b), which are luggage and backpack respectively. Taking the above images as examples, visualization studies were conducted on CBAM residual blocks 2-2, CBAM residual blocks 3-2, and CBAM residual blocks 4-4. The output weight of the spatial attention module and channel attention module are shown in figure 8 and figure 9 respectively, where white represents an output weight of 1 and black represents an output weight of 0. It can be seen from figure 8 that corresponding to the blue part of the X-ray image, the spatial attention module outputs a greater weight. The spatial attention module focuses more on the interior of the package, especially the blue part, so as to improve the network's feature expression ability. It can be seen from figure 9 that the channel attention module focuses more on features with sharp contrast and more prominent wrapping contours, so as to improve the network's feature expression ability.

Figure 7. (a) X-ray image of security inspector

Figure 7. (b) X-ray image of security inspector

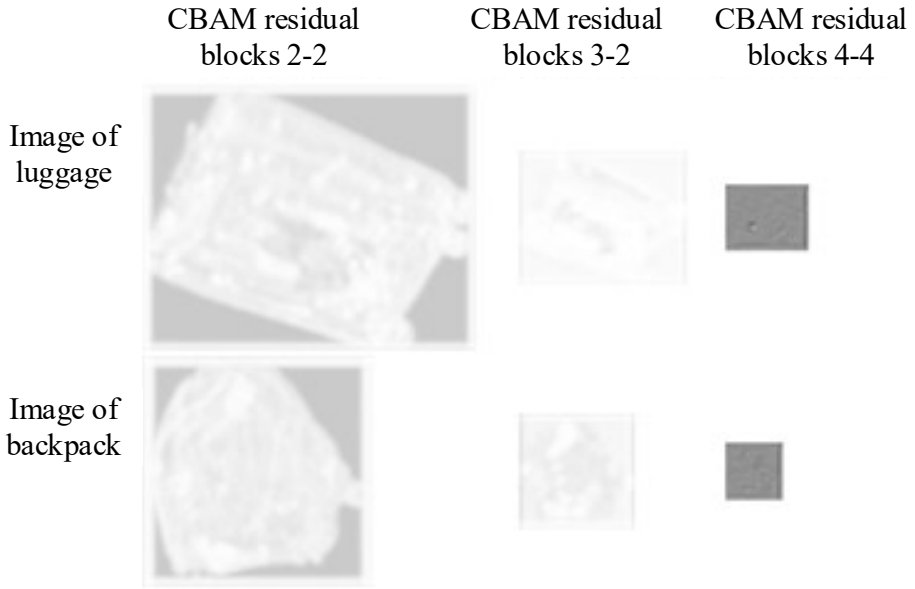

Figure 8. Output weight of spatial attention module

DOI: 10.54254/2755-2721/75/20240552

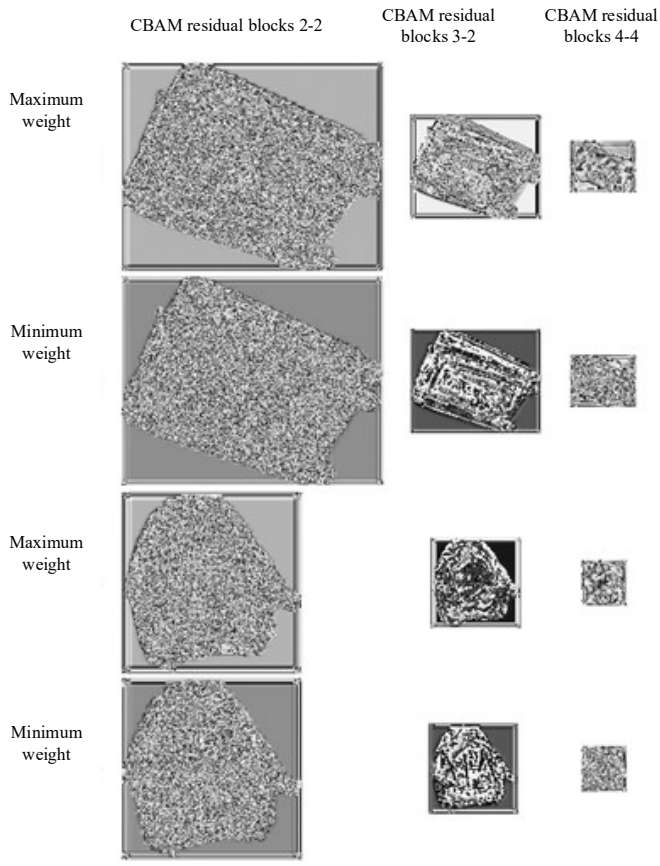

Figure 9. Output weight of channel attention module

4.3.4. Performance comparison of the improving methods

The detection performance of the proposed method was compared with the original CenterNet method and other algorithms. The experimental results are shown in table 3. The accuracy of the improved CenterNet increased by 3.3% compared to that of the original CenterNet, with little reduction in detection speed. With higher detection speed, the accuracy increased by 5.5%, 2.4%, 3.6%, and 10.1% compared to that of SSD, Yolo V3, FPN, and DERT[20] algorithms. The detection accuracy of contraband in X-ray images is effectively improved.

Table 3. Test results of different models

Algorithm	Gun	Knife	Scissor	Wrench	Plier	mAP	FPS
SSD	90.4	75.8	83.1	78.7	83.9	82.4	30.6
Yolo-v3	90.5	78.0	89.3	79.9	89.6	85.5	45.3
FPN(ResNet50)	90.7	79.7	80.9	83.0	87.0	84.3	21.9
Dert(ResNet50)	88.3	77.0	77.5	68.5	77.6	77.8	14.7
CenterNet	90.7	80.0	83.9	78.9	89.4	84.6	59.2
Improved CenterNet	90.8	84.4	89.5	85.9	89.3	87.9	46.7

Ablation experiments were conducted on the proposed method in this article to analyze the impact of each module. The experimental results are shown in table 4. The improved ResNet18, CBAM module, and adaptive feature fusion all increase detection accuracy, and the highest accuracy is achieved when all three are combined, with a 3.3% increase in accuracy compared to the original CenterNet network.

DOI: 10.54254/2755-2721/75/20240552

Table 4. Results of ablation experiment

Combination mode	Improved ResNet18	CBAM module	Adaptive feature fusion	mAP
Combination 1				84.6
Combination 2	√			86.6
Combination 3	√	√		87.6
Combination 4	√		√	87.4
Combination 5	√	√	√	87.9

5. Conclusion

An improved CenterNet method is proposed, and an improved ResNet18 network is designed based on analysis of the imaging mechanism and target characteristics of X-ray images. An attention mechanism is introduced, and an adaptive feature fusion method is designed to improve the accuracy of contraband detection in X-ray images. Experiments were conducted on the SIXray dataset. The output features of convolutional layers with different down-sampling rates are demonstrated through feature visualization, revealing the mechanisms of spatial attention and channel attention of X-ray images. The experiments show that: (1) The backbone network with a down sampling rate of 16 times is more suitable for detecting contraband in X-ray images. (2) The channel attention module enhances the network's feature expression ability by focusing more on features of sharp contrast and more prominent wrapping contours, while the spatial attention module improves the network's feature expression ability by focusing more on the interior of the package, especially the blue part. (3) The improved CenterNet network increased accuracy by 3.3% over the original CenterNet network by integrating three methods: 16 times down sampling rate backbone network, CBAM module, and adaptive feature fusion. This method is more accurate than one-stage and two-stage target detection methods. The detection accuracy of contraband in X-ray images is effectively improved. In the next step, incorporating the railway passenger security inspection scenario, a dataset of contraband in X-ray images with more types and quantities will be constructed to verify the effectiveness of the algorithm.

Acknowledgements

This paper was financially supported by two projects. One is “Research on Railway Passenger Transport Security Inspection Management System and Centralized Image Judgement Operation Mode” (N2023X004) by National Railway Administration of the People's Republic of China. The other one is "Image generation technology for prohibited and restricted items in passenger transportation safety inspection" (DZYF23-32) by Beijing Jingwei Information Technology Co., Ltd.

References

[1] ZHANG Q L, TANG W, YANG D. Railway Passenger Transport Safety Inspection Management Information System Based on Intelligent Identification Technology[C]. //Proceedings of the 16th China Intelligent Transportation Annual Conference. 2021:264-271.

[2] HE N, ZHANG Z L, MA W D, et al.Construction of Smart Safety Inspection Sys-tem in Urban Rail Transit[J]. Urban Mass Transit, 2022,25(04):214-216+220.

[3] CHANG Q Q, CHEN J M, LI W J, et al. Dangerous Goods Detection Technology Based on X-ray Images in Urban Rail Transit Security Inspection[J].Urban Mass Transit, 2022,25(04):205-209.

[4] Miao C , Xie L , Wan F , et al. SIXray: A Large-Scale Security Inspection X-Ray Benchmark for Prohibited Item Discovery in Overlapping Images[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.

[5] Ma C, Zhuo L, Li J, et al. Prohibited Object Detection in X-ray Images with Dynamic Deformable Convolution and Adaptive IoU[C]//2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022: 3001-3005.

DOI: 10.54254/2755-2721/75/20240552

[6] Kumar R S, Balaji A, Singh G, et al. Recursive CNN Model to Detect Anomaly Detection in X-Ray Security Image[C]//2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM). IEEE, 2022, 2: 742-747.
[7] Gaus Y F A , Bhowmik N , Breckon T P . On the Use of Deep Learning for the Detection of Firearms in X-ray Baggage Security Imagery[C]// IEEE Symposium on Technologies for Homeland Security (HST 2019). IEEE, 2019.
[8] Wei Y, Dai C, Chen M, et al. Prohibited Items Detection in X-ray Images in YOLO Network[C]//2021 26th International Conference on Automation and Computing (ICAC). IEEE, 2021: 1-6.
[9] ZHU C, LI B Y, LIU X Q, et al. A deep convolutional neural network based on YOLO for contraband detection[J]. Journal of Hefei University of Technology(Natural Science) ,2021,44(9):1198-1203.
[10] Song B, Li R, Pan X, et al. Improved YOLOv5 Detection Algorithm of Contraband in X-ray Security Inspection Image[C]//2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI). IEEE, 2022: 169-174.
[11] Wang Z, Zhang H, Lin Z, et al. Prohibited Items Detection in Baggage Security Based on Improved YOLOv5[C]//2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI). IEEE, 2022: 20-25.
[12] ZHANG H, ZHANG S C. Security Inspection Image Object Detection Method with Attention Mechanism and Multilayer Feature Fusion Strategy[J]. Laser & Optoelectronics Progress, 2022,59(16):197-208.
[13] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. Springer, Cham, 2014: 818-833.
[14] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
[15] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
[16] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.
[17] Liu S, Huang D, Wang Y. Learning spatial fusion for single-shot object detection[J]. arXiv preprint arXiv:1911.09516, 2019.
[18] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
[19] Chen K, Wang J, Pang J, et al. MMDetection: Open mmlab detection toolbox and benchmark[J]. arXiv preprint arXiv:1906.07155, 2019.
[20] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv:2010.04159, 2020.