Proceedings of CECNet 2022
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220521

PIXDet: Prohibited Items X-Ray Image Detection in Complex Background

Mingyuan LI [a], Bowen MA [a], Tong JIA [a,1] and Yichun ZHANG [b]
[a] *College of Information Science and Engineering, Northeastern University, Shenyang, China.*
[b] *China Institute of Arts Science & Technology, Beijing, China*

Abstract. In this paper, aiming at the complex background and overlapping characteristics in X-ray images, we propose an unique spatial attention mechanism based on the feedback of high-level semantic feature to guide low-level semantic features, named Feedback Guidance Mechanism (FGM). In addition, in view of the high probability of miss of small prohibited items, a feature aggregation method based on the fusion of high and low-level features and dilated convolution is proposed, named Feature Aggregation Module (FAM). Then, we combine FGM and FAM into a lightweight model SSD and get a new Prohibited Items Detector (PIXDet). Our experiments indicate that PIXDet is more lightweight, but it can achieve 90.36% mAP on PIXray dataset, exceeding SSD by 1.0% mAP, outperforming some state-of-the-art methods, implying its potential applications in prohibited items detection field.

Keywords. Deep Learning, security inspection, attention mechanism, feature aggregation

1. Introduction

With the growth of population in large cities and crowd density in public transportation hubs, security inspection has been playing an increasingly critical role in maintaining urban public safety[1]. In densely populated places such as stations or airports, the use of X-ray scanners to assist staff in checking passenger packages is one of the most extensive and effective security detection methods. The various objects in the detection image overlap each other with different shapes. It is difficult to find the prohibited items hidden in them, especially after the security inspectors have worked for a long time. With the development of deep learning technology, more and more research begin to use computer-aided technology based on object detection technology to assist human inspectors in security inspection[1,2,3,4,5]. Recently, some works using convolutional neural networks to detect the prohibited items in X-ray image have achieved rich results[1,4,5].

However, X-ray baggage images where objects are randomly stacked and heavily overlapped each other, resulting that background interfere with foreground detection and small prohibited items are easily missed. These characteristics bring great challenges to

[1]Corresponding Author: Tong Jia, a Professor with the College of Information Science and Engineering, Northeastern University, Shenyang, China. E-mail: jiatong@ise.neu.edu.cn

both object detection methods and human inspectors[1]. We mainly focuses on the detection difficulties in X-ray security images, and propose a network framework based on Single Shot MultiBox Detector (SSD) algorithm[6] to adapt to prohibited items detection in X-ray images —PIXDet. The main contributions are as follows:

1) A new attention mechanism (FGM) is proposed, which uses high-level features to guide low-level features fusions in order to remove background information. By simulating the process of continuous feedback iteration, the network will pay more attention to the target area and remove redundant information. FGM effectively solves the problem of background interference in X-ray images, enables the network to efficiently and accurately detect prohibited items in X-ray images with complex backgrounds.

2) A feature aggregation method (FAM) is proposed to strengthen the expression of detailed features, integrate contextual information, and effectively improve the problem of missed detection of small prohibited items. It has a strong adaptability for contraband with different aspect ratio and scale change.

3) We combine SSD, FGM and FAM into a Prohibited Items Detector, named PIXDet. After a series of experiments and analysis, we come to the conclusion that our model is more lightweight while outperforming the state-of-the-art methods, implying its potential applications in prohibited items detection fields.

2. Related Work

2.1. Attention Mechanism

Attention mechanism can be understood as a kind of algorithm that helps computer to put the more computational resources into the most informative components of a signal[5]. Recently, there have been a series of studies incorporating attention mechanism improve the performance of computer vision tasks, including image classification, object detection and semantic segmentation[7,8,9,10]. SE-Net[9] proposes an effective attention mechanism utilizing channel attention for the first time and obtains promising results. But, it does not utilize spatial information, CBAM[10] based on SE-Net employs max pooling and average pooling to extract spatial feature. However, the pooling operations will result that the feature extraction network mix information of foreground and background. We use FGM solves this problem to some extent.

2.2. Methods of Feature Aggregation

FPN[11] is one of the most famous methods improve detection and segmentation performance by using features from different layers. PANet[12] proposes one bottom-up path augmentation mechanism to shorten information path and enhance FPN by more accurate localization signals in low-level feature. SSD[6], FSSD[13] and DSSD[14] utilize different feature levels to inference proposals. OPIXray[5] refines feature maps by operation named RIA for extracting more useful information from image better. However, RIA just uses the same size of convolution kernel, resulting that the feature extraction network ignores the information of the small objects in image. We propose a FAM mechanism solves this problem to some extent.

2.3. Object Detection Studies for Security Inspection

With the vigorous improvement of computer computing power and convolutional neural network technology, a large number of object detection algorithms emerge in an endless stream, i.e., Fast R-CNN[7], Faster R-CNN[8], SSD[6], DSSD[14], FSSD[13], YOLO[15], YOLOv3[16] and YOLOv4[17]. SSD is one of the most famous neural networks which is fast and high precision and our work is based on it. We boost it with FGM and FAM which are proposed according to the characteristics of X-ray image of security inspection. To the best of our knowledge about X-ray datasets of security inspection, there only three X-ray datasets are open for research purposes, i.e., SIXray[1], OPIXray[5] and PIXray[4]. SIXray[1] publics a large-scale X-ray dataset including 105,931 X-ray images, but it just consists of six categories of prohibited items. OPIXray[5] just contains five kinds of knives, including folding knife, straight knife, scissor, utility knife, and multi-tool knife, lacking of resemblance to the real world. Therefore, we choose PIXray dataset which contains fifteen kinds of prohibited items and 5,046 X-ray images, in which 15 classes of 15,201 prohibited items[4].

3. Proposed Method

Challenge against the above, based on the single-stage object detector SSD network model, we propose a feedback guidance mechanism (FGM), which uses high-level semantic features to guide low-level semantic features after being processed by some modules, improving the feature representation of the region of interest, and thus realizing more accurate identification and positioning of overlapping prohibited items. In addition, in order to improve the detection of small prohibited items, feature aggregation module (FAM) is added to capture the lost information of local small targets in the global field of vision and reduce the missed detection rate of small prohibited items. Finally, we propose the PIXDet to combine the two modules and SSD to achieve the best result. The overall architecture is shown in Figure 1. In the following subsections, we first introduce the overall PIXDet, then elaborate on the proposed FGM, and finally describe FAM.

3.1. Overall Architecture

In fact, our PIXDet is independent of backbone, it can be easily applied to mainstream backbone networks, i.e., vgg-16[18], resnet-50[19] and densenet-121[20]. We just use the lightest vgg networks as backbone for validating our theory, which is the same as primary SSD[6]. The dataset $X = \{x_1, ..., x_N\}$ has N training images, in Figure 1, each input image $x_n \in X$ will be fed into backbone of PIXDet to extract L different feature layers $x_n^{(l)}$, l represents the l-th of feature layer which will be directly used to detection module in primary SSD[6]. The larger the number of l, the higher the feature layer. One $FGM^{(l)}$ will be fed into two feature layers $x_n^{(l)}$ and $x_n^{(l+1)}$, outputting $\tilde{x}_n^{(l)}$, and the light green area in the Figure is where FGM plays a role.

Then, the guided feature layers $\tilde{x}_n^{(l)}$ and $\tilde{x}_n^{(2)}$ will be fused into $\hat{x}^{(1)}$ by FAM, and the light orange area in the Figure 1 is where FAM plays a role. Later, similar to SSD, we feed $\hat{x}_n^{(1)}, \tilde{x}_n^{(2)}, ..., \tilde{x}_n^{(L)}$ to prediction layers to get classification and regression information, where L is 6 in our model. Finally, we paint the decode information predicted by PIXDet

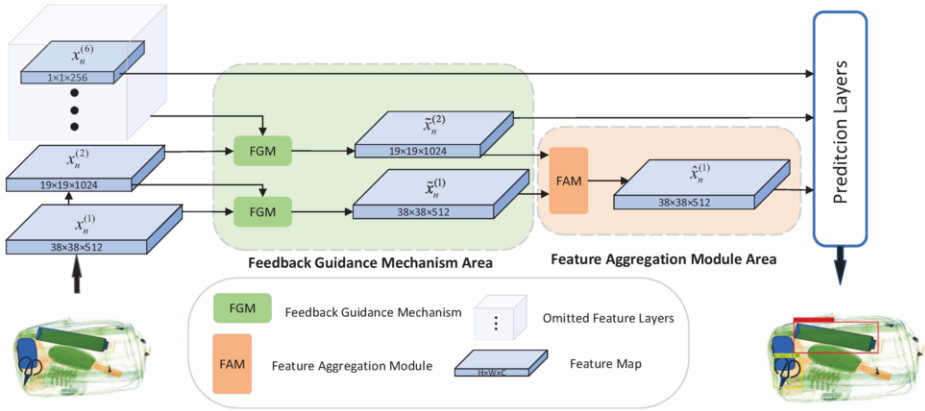

Figure 1. Prohibited Items X-ray Detector. The network adds a spatial attention mechanism of high-level semantic feature feedback to guide low-level semantic features and feature aggregation methods that combine high and low level features with dilated convolution.

on the input image to get output image. Summarizing PIXDet yields the following optimization problem:

$$\alpha^*, \beta^*, \gamma^* = arg \min_{\alpha,\beta,\gamma} \sum_{n=1}^{N} Loss(y_n, y_n^*) \tag{1}$$

$$y_n = f(\hat{x}_n^{(1)}, \tilde{x}_n^{(2)}, ..., \tilde{x}_n^{(L)}; \gamma) \tag{2}$$

$$\tilde{x}_n^{(l)} = FGM^{(l)}(x_n^{(l)}, x_n^{(l+1)}; \alpha^{(l)}) \tag{3}$$

$$\hat{x}_n^{(1)} = FAM^{(1)}(\tilde{x}_n^{(1)}, \tilde{x}_n^{(2)}; \beta^{(1)}) \tag{4}$$

where $Loss(\cdot,\cdot)$ is a loss function which is the same as SSD[6]. Within this process, $\alpha^{(l)}$ and $\beta^{(1)}$ are parameters in $FGM^{(l)}$ and $FAM^{(1)}$, which will be updated according to ground-truth y_n^*, while γ is the other parameters automatically updated according to back propagation in model. $FGM^{(l)}(\cdot,\cdot)$ and $FAM^{(l)}(\cdot,\cdot)$ are the simplified equations of FGM and FAM, which are respectively discussed in Section 3.2 and Section 3.3.

3.2. Feedback Guidance Mechanism (FGM)

As shown in Figure 2, the FGM consists of a CR_{-1} and an up-sampling layer with bilinear interpolation, and then uses the SoftMax activation function for each channel to obtain a probability diagram for each channel. CR_{-1} is a module including convolution and Relu, and the kernel size and stride is 1. Specifically, the feature maps of Conv4_3 layer and FC7 layer are recorded as $x_n^{(1)}$ and $x_n^{(2)}$, respectively, where $x_n^{(1)} \in \mathbb{R}^{512\times38\times38}$,$x_n^{(2)} \in \mathbb{R}^{1024\times19\times19}$. We get the intermediate guidance weight chart to $\dot{x}_n^{(2)}$, where $\dot{x}_n^{(2)} \in \mathbb{R}^{512\times38\times38}$. We get $\dot{x}_n^{(2)}$ by calculating as follows:

$$\dot{x}_n^{(2)} = Relu(Conv_{1\times1}(x_n^{(l)})) \tag{5}$$

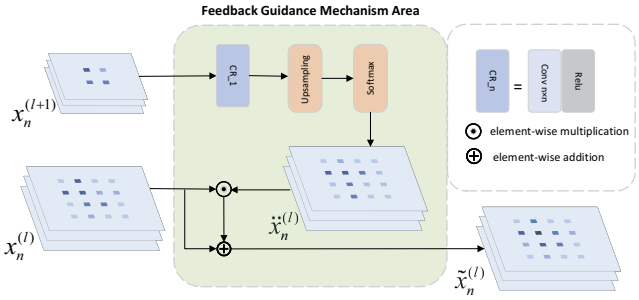

Figure 2. Feedback Guidance Mechanism: Higher-level semantic features are superimposed with the intermediate guidance weight $\dot{x}_n^{(2)}$ to low-level semantic features.

$$\ddot{x}_n^{(2)} = SoftMax(UpSampling(\dot{x}_n^{(2)}) \tag{6}$$

In order to obtain the same number of channels as $x_n^{(1)}$, $x_n^{(2)}$ is passed through a 1×1 convolution operation and activated by Relu, getting $\dot{x}_n^{(2)} \in \mathbb{R}^{512\times19\times19}$. Then it will be upsampled by bilinear interpolation for changing feature map dimension, while the SoftMax function is used to normalize the feature values in order to optimize the weight of all channel feature expressions and strengthen the feature expression of contraband regions in low-level semantic features.

Finally, the intermediate guidance weight $\ddot{x}_n^{(2)}$ is fused with $x_n^{(1)}$, which serves as a correction. The fusion process can be expressed as:

$$\tilde{x}_n^{(1)} = \varepsilon^{(1)}(x_n^{(1)} \odot \ddot{x}_n^{(2)}) \oplus x_n^{(1)} \tag{7}$$

among them, $\oplus$ indicates element-wise addition, $\odot$ denotes element-wise multiplication, and $\varepsilon^{(1)}$ is an adaptive learning factor for 1st layer, which is used to auto-modify the power of influence of the higher-lever feature layer. Eqs. (5,6,7) can be simplified to Eq. (3).

Note that the FGM is developed in a unified framework that allows calculations to be trained end-to-end through back propagation of all layers. We just choose two layers to show its detail in Section 3.2, each FGM is not completely consistent due to the different dimensions of the data processed. It elegantly handles the main limitations of the existing one-stage object detector applied to prohibited items X-ray image detection. The proposed feedback guidance mechanism can effectively improve the interference of complex background in contraband detection, and at the same time allow the X-ray images to be varied, which further improves the performance of the detection.

3.3. Feature Aggregation Module (FAM)

Low-level features have a small receptive field, which can provide more detailed information for the recognition of small prohibited items (such as razor blades), but low-level features have no semantic feature information, and with the deepening of the network, the local small target information under the global vision will also be lost, which is the main reason for the high rate of missed detection of small targets. Therefore, we propose a feature aggregation module (FAM), whose structure is shown in Figure 3.

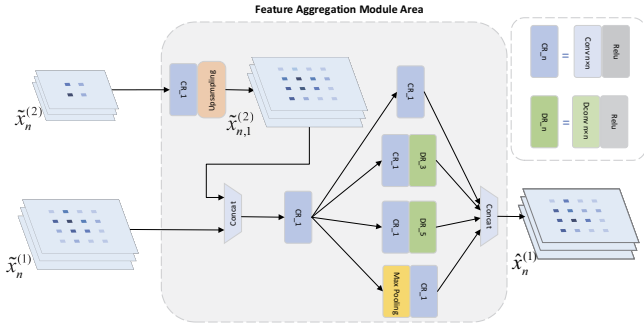

Figure 3. Feature Aggregation Module: Strengthen the expression of details and features and integrate contextual information.

Firstly, the higher level semantic feature FC7 layer $\tilde{x}_n^{(2)} \in \mathbb{R}^{1024\times19\times19}$, which has been processed by $FGM^{(1)}$ will be upsampled and convolved as the operation of Eq. (8):

$$\tilde{x}_{n,1}^{(2)} = UpSampling(Relu(Conv_{1\times1}(x_n^{(1)}))) \tag{8}$$

The obtained feature map of Eq. (8) is denoted as $\tilde{x}_{n,1}^{(2)} \in \mathbb{R}^{512\times38\times38}$, while Conv4_3 layer $\tilde{x}_n^{(2)} \in \mathbb{R}^{512\times38\times38}$ will be fused with $\tilde{x}_{n,1}^{(2)}$ to get the transitional feature map $\tilde{x}_{n,2}^{(2)}$:

$$\tilde{x}_{n,2}^{(2)} = Relu(Conv_{1\times1}(\tilde{x}_n^{(1)}||\tilde{x}_{n,1}^{(2)})) \tag{9}$$

where $||$ represents a concatenate operation, $\tilde{x}_n^{(1)}||\tilde{x}_{n,1}^{(2)} \in \mathbb{R}^{1024\times38\times38}$. We use 1×1 convolution kernels for dimensional reduction to obtain the associated feature map $\tilde{x}_{n,2}^{(2)} \in \mathbb{R}^{512\times38\times38}$.

In addition, dilated convolution[21] can increase the receptive field of convolution kernel while keeping the number of parameters unchanged, which is conducive to output more global feature information and help other algorithms learn global semantic feature information. At the same time, inspired by the inception architecture in GoogleNet[22], dilated convolution is added into the inception module to expand the visual field free of charge, which allows FAM to use multi-scale visual field to establish global and local connection to take into account the task of global positioning and local classification, effectively improving the detection rate of small prohibited items .

The latter half of the FAM is processed as follows:

$$\tilde{x}_{n,3}^{(1)} = Relu(Conv_{1\times1}(\tilde{x}_{n,2}^{(1)})) \tag{10}$$

$$\tilde{x}_{n,4}^{(1)} = Relu(Dconv_{3\times3}(Relu(Conv_{1\times1}(\tilde{x}_{n,2}^{(1)})))) \tag{11}$$

$$\tilde{x}_{n,5}^{(1)} = Relu(Dconv_{5\times5}(Relu(Conv_{1\times1}(\tilde{x}_{n,2}^{(1)})))) \tag{12}$$

$$\tilde{x}_{n,6}^{(1)} = Relu(Conv_{1\times1}(MaxPooling_{3\times3}(\tilde{x}_{n,2}^{(1)}))) \tag{13}$$

$$\hat{x}_n^{(1)} = \tilde{x}_{n,3}^{(1)}||\tilde{x}_{n,4}^{(1)}||\tilde{x}_{n,5}^{(1)}||\tilde{x}_{n,6}^{(1)} \tag{14}$$

where $\hat{x}_n^{(1)} \in \mathbb{R}^{512\times38\times38}$, $\tilde{x}_{n,3}^{(1)} \in \mathbb{R}^{128\times38\times38}$, $\tilde{x}_{n,4}^{(1)} \in \mathbb{R}^{256\times38\times38}$, $\tilde{x}_{n,5}^{(1)} \in \mathbb{R}^{64\times38\times38}$ and $\tilde{x}_{n,6}^{(1)} \in \mathbb{R}^{64\times38\times38}$. DR_n is a module including dilated convolution and Relu, and the kernel size and dilation of dilated convolution is 'n' and 2 respectively. $FAM^{(1)}$ including Eqs. (8-14) can be simplified to Eq. (4) in section 3.1. Note that our model just has one FAM named $FAM^{(1)}$, '1' in equation represents the feature layers location of it.

4. Experiments

In this section, we carry on extensive experiments to evaluate FGM and FAM we propose. Firstly, the superiority in detecting prohibited items in X-ray images of PIXDet is verified through some comparative experiments. Secondly, the ablation experiment of the model is carried out for illustrating completely the effectiveness of each module we propose above. The experimental results demonstrate that the method proposed in this paper achieves satisfying results in the detection of contraband in X-ray images. Finally, we perform feature map visualization to show the effect of FGM. In the following experiments, the hardware and software environments of the experiment are PyTorch, Windows 10 system, and NVIDIA 3070 laptop GPU. All models are optimized by the SGD optimizer and the learning rate is set to 0.0001. The batch size is set to 8, and momentum decay and weight decay are set to 0.9 and 0.0005, respectively. What's more, we do not use pre-trained model of Imagenet 1000, because our detection task is based on PIXray dataset which is independent of natural images. For fairness, all models are trained for 300 epochs with the same hyper-parameter and training strategy.

4.1. Comparisons with State-of-the-art Methods

Taking SSD[6] as the basic architecture, in order to prove the effectiveness of the design of functional modules, we compare our PIXDet with some SOTA object detection models including SSD[6], YOLOv3[16] and FSSD[13] on the dataset PIXray with input size 300×300. The ratio of training set and test set is 9:1 and mean Average Precision (mAP) was used as the index to evaluate the detection accuracy. The detection accuracy of prohibited items of X-ray images in each model is shown in Table 1. We evaluate the mean Average Precision (mAP) of the object detection to measure the performance of the model and the IOU threshold is set to 0.5.

Table 1. Complexity comparison of different models. PARAMs, SIZE and GFLOPs represent the total number of parameters, the model size and the Giga Floating Point operations, respectively.

Method	mAP	PARAMs	GFLOPs	SIZE(MB)
SSD[6]	89.36	**25.48**×10^6	31.10	**97.7**
YOLOv3[16]	85.71	61.59×10^6	**19.42**	235
FSSD[13]	90.20	33.05×10^6	37.27	136
PIXDet(ours)	**90.36**	27.83×10^6	33.03	106

PIXDet achieves 90.36% mAP on PIXray dataset with input size 300×300, outperforming SSD, YOLOv3 and FSSD by 1.0% mAP, 4.65% mAP and 0.16% mAP, respectively. With regard to PARAMs and SIZE, PIXDet is only second to SSD. Our method

only brings a slight increase in computational cost(6.2% in GFLOPs), compared with SSD without any improvement. In brief, according to the experimental results, mAP of PIXDet is better than other state-of-the-art models. Both GFLOPs and model SIZE are lower than FSSD, meaning that it requires lower calculation power and can be more easily developed on embedded devices. However, comparing with SSD[4], there is a small increase in the demand for computing resources, i.e., the number of parameters, model size and the need of float point operation.

4.2. Ablation study

In this part, an ablation experiment based on SSD model is designed, which could more intuitively show the influence of FGM and FAM on the detection effect of PIXDET. The experimental parameters are consistent with the contrast experiment. The results of ablation experiment are shown in Table 2.

Table 2. Ablation studies of PIXDet. "+FGM" represents adding feedback guidance mechanism to SSD, "+FAM" represents adding feature aggregation module to SSD.

Method	mAP	Bat	Lighter	Pliers	Gun	PV	Scis	Hamm	Wren	Batt	Fires	Dart	Saw	Screw	Knife	Razor
vanilla model	89.36	100	96	98	96	96	97	95	93	90	**91**	76	**82**	80	82	69
+FGM	89.72	100	96	98	97	94	**97**	**96**	**95**	92	89	82	77	**81**	76	78
+FAM	90.05	100	96	96	**98**	**97**	96	95	94	**94**	87	83	79	78	82	77
+FGM+FAM(PIXDet)	**90.36**	**100**	**98**	**98**	96	95	94	93	94	92	90	**83**	78	80	**85**	**79**

FGM and FAM can improve detection performance by 0.36% and 0.71% on a basic detection network (SSD), respectively. It is concluded that when the detection network increases the corresponding feature weight through reverse connection, FGM is capable of completing the optimization of feature information adaptively, focusing the attention of the model more on prohibited items, reducing the interference degree of middle-level useless information and improving the detection accuracy. Compared to some other types of prohibited items, it is particularly difficult to detect the knife due to its small size. FAM has effectively improved the detection accuracy of small prohibited items, and the detection accuracy of dart has increased from 76% to 83%. PIXDet means adding module FGM and FAM on SSD. It can be seen that FGM and FAM are complementary to each other to achieve better performance. Therefore, mAP of PIXDet is the best of them. It is worth mentioning that although some kinds of prohibited items can not be detected perfectly by PIXDet, the detection results of them are not much worse than the vanilla model. In brief, the experimental results show that PIXDet has certain effectiveness in prohibited items object detection. However, PIXDet is not without its shortcomings, comparing with '+FGM' or '+FAM', the average precision of PIXDet has a slight drop, meaning that there is some overlap between FGM and FAM. Even so, they are functionally complementary because the mean average precision of PIXDet is higher than '+FGM' and '+FAM'.

4.3. Visualization Analysis

We display the detection results of PIXDet, as shown in Figure 4. SSD can only detect some characteristic objects, i.e., bat and pressure vessel. When it comes to small prohibited items razor blade and battery, the rate of missing or false detection is extremely

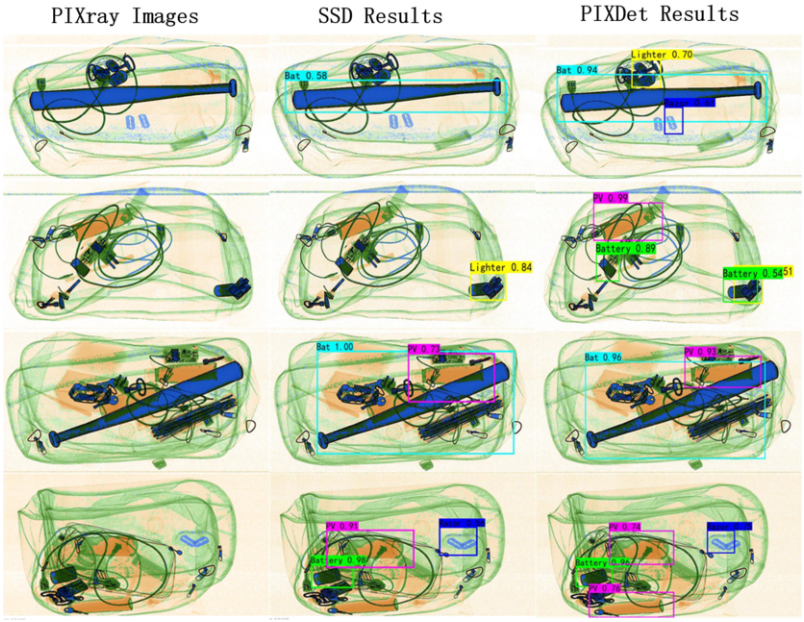

Figure 4. Detection examples using baseline model(SSD) and our model.

high. In contrast, our PIXDet can adapt to the X-ray images with complex background, and it can even recognize battery overlapped with electric wire. By comparing the experiments results, it can be demonstrated that our model can effectively improve the problems of background interference and small contraband missed detection in X-ray security images.

5. Conclusion

In this work, we explore in detail the application of deep learning in the object detection task of prohibited items in X-ray images. To facilitate research in this area, we propose the PIXDet detection model to solve the existing problems in contraband images, such as complex background and missed detection of small objects. Specifically, feedback guidance mechanism (FGM) and feature fusion module (FAM) proposed in this paper can effectively weaken the interference caused by complex background and improve the detection accuracy of small prohibited items. We have comprehensively evaluated our detection model on PIXray dataset, and proved that our module can effectively improve the detection performance and provided a new idea for the model improvement. Furthermore, our model is more lightweight while outperforming the state-of-the-art methods, implying its potential applications in prohibited items detection field. In the future, we will continue to study object detection of smaller prohibited items in more complex and real background.

References

[1] 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]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2119-2128.

[2] Griffin L D, Caldwell M, Andrews J T A, et al. "Unexpected item in the bagging area": anomaly detection in X-ray security images[J]. IEEE Transactions on Information Forensics and Security, 2018, 14(6): 1539-1553.

[3] Abidi B R, Zheng Y, Gribok A V, et al. Improving weapon detection in single energy X-ray images through pseudocoloring[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2006, 36(6): 784-796.

[4] Ma B, Jia T, Su M, et al. Automated Segmentation of Prohibited Items in X-ray Baggage Images Using Dense De-overlap Attention Snake[J]. IEEE Transactions on Multimedia, 2022.

[5] Wei Y, Tao R, Wu Z, et al. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 138-146.

[6] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[7] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[8] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[9] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[10] 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.

[11] 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.

[12] 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.

[13] Li Z, Zhou F. FSSD: feature fusion single shot multibox detector[J]. arXiv preprint arXiv:1712.00960, 2017.

[14] Fu C Y, Liu W, Ranga A, et al. Dssd: Deconvolutional single shot detector[J]. arXiv preprint arXiv:1701.06659, 2017.

[15] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[16] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[17] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[19] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[20] Iandola F, Moskewicz M, Karayev S, et al. Densenet: Implementing efficient convnet descriptor pyramids[J]. arXiv preprint arXiv:1404.1869, 2014.

[21] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.

[22] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.