Pattern Recognition 122 (2022) 108261

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Exploiting foreground and background separation for prohibited item detection in overlapping X-Ray images

Fangtao Shao, Jing Liu*, Peng Wu, Zhiwei Yang, Zhaoyang Wu

Guangzhou Institute of Technology, Xidian University, Guangzhou, China

ARTICLE INFO

Article history:
Received 3 January 2021
Revised 9 August 2021
Accepted 18 August 2021
Available online 20 August 2021

Keywords:
X-ray imagery
Object detection
Foreground and background separation (FBS)
Recursive training

ABSTRACT

X-ray imagery security screening is an essential component of transportation and logistics. In recent years, some researchers have used computer vision algorithms to replace inefficient and tedious manual baggage inspection. However, X-ray images are complicated, and objects overlap with one another in a semi-transparent state, which underperforms the existing object detection frameworks. To solve the severe overlapping problem of X-ray images, we propose a foreground and background separation (FBS) X-ray prohibited item detection framework, which separates prohibited items from other items to exclude irrelevant information. First, we design a target foreground and use recursive training to adaptively approximate the real foreground. Thereafter, with the constraints of X-ray imaging characteristics, a decoder is employed to separate the prohibited items from other irrelevant items to obtain the foreground and background (FB). Finally, we use the attention module to make the detection framework focus more on the foreground. Our method is evaluated on a synthetic dataset with FB ground truth and two public datasets with only bounding box annotations. Extensive experimental results demonstrate that our method significantly outperforms state-of-the-art solutions. Furthermore, experiments are performed in the case where only a small number of images contain the FB ground truth. The results indicate that our method requires only a small number of FB ground truths to obtain a performance equivalent to that of all FB ground truths.

1. Introduction

X-ray imagery security screening is widely used to maintain transportation and logistics security, effectively reducing the risk of crime and terrorist attacks. Manual screening by human operators remains the primary screening method. However, inspection is a complex task because prohibited items are always occluded by other objects or present in various illegible views and only account for a small portion of actual X-ray images. In addition, human inspectors only have a few seconds to decide whether a bag contains prohibited items during rush hour. Hence, an algorithm that can automatically and accurately identify prohibited items in X-ray imagery has become increasingly important.

In contrast to photographic (optical) imagery, which is formed by light reflecting an object, X-ray imagery is formed by irradiating an object with X-rays that pass through the object. According to the absorption law, objects of different structures and densities attenuate X-rays differently [1]. Therefore, objects in an X-ray image heavily overlap in a translucent state and lose texture and gloss characteristics (as shown in Fig. 1), making the object detection algorithm for general photographic images underperform on X-ray images.

In recent years, deep learning has achieved remarkable results in image, audio [2], and videos [3,4] detection. Thus, many researchers have applied deep learning methods to X-ray scenes in photographic scenes. Li *et al.* [5] applied multitask contrastive learning for automatic X-ray diagnosis. Wang *et al.* [6] proposed a novel framework for differentiating and localising COVID-19 from community-acquired pneumonia. In addition to medical X-ray applications, deep learning is widely used in X-ray security screening. Akçay *et al.* [7] trained convolutional neural network (CNN)-based techniques, such as the faster region-based CNN (Faster RCNN) [8] and You Only Look Once version 2 (YOLOv2) [9] models on the firearm detection problem under X-ray conditions. Liang *et al.* [10] evaluated the performance of Fast RCNN and single-shot multibox detector (SSD) [11] using single/multiview X-ray imagery. Gaus *et al.* [12] employed a transfer learning approach to evaluate whether such inter-scanner generalisations may exist over a multiple-class detection problem. These studies show that com-

* Corresponding author.
E-mail addresses: shaofangtao96@163.com (F. Shao), neouma@163.com (J. Liu), xdwupeng@gmail.com (P. Wu), zwyang97@163.com (Z. Yang), 15191737495@163.com (Z. Wu).

https://doi.org/10.1016/j.patcog.2021.108261
0031-3203/

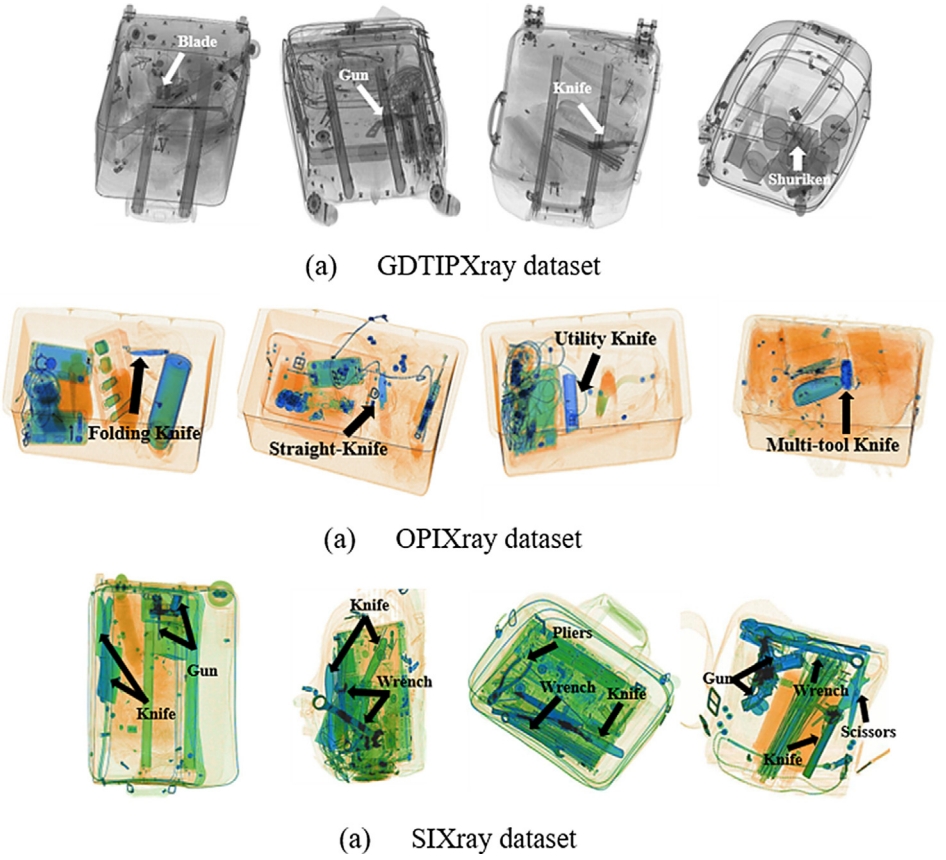

Fig. 1. Examples of X-ray images showing severely overlapping items that lack texture and gloss characteristics: (a), (b), and (c) are obtained from the GDTIPXray, OPIXray, and SIXray datasets, respectively.

mon object detection methods have certain applicability to X-ray scenes, but they did not improve the algorithm by making use of the unique properties of X-rays.

Furthermore, some researchers have begun to study the problem of complex items and serious occlusion in X-rays. Miao *et al.* proposed a class-balanced hierarchical refinement (CHR) model [13]. To eliminate irrelevant information, generative models and information from different neighbouring scales were used. However, preserving objects' locality information among different scales is extremely difficult, especially when they are non-coherent. Because the edges and contours of an image can sufficiently reflect its characteristics [14], Wei *et al.* proposed the de-occlusion attention module (DOAM) [15], which uses two sub-modules to emphasise the edge information and material information of an object. However, in the edge-enhanced image obtained by the Sobel operator, the edge information of prohibited items and irrelevant items are still mixed, and the network still needs to identify the prohibited items from the mixed edge information. In [16], Liu *et al.* used a colour distribution to separate the foreground and background (FB). They defined an RGB threshold, where the part of the image within the threshold will be regarded as the background or the foreground. Although most of the prohibited items have been successfully classified into the foreground, irrelevant items of similar colours may still overlap with prohibited items due to this setting.

In this paper, we propose a foreground and background separation (FBS) method for X-ray object detection, which can not only detect prohibited items but also obtain pure images of prohibited items from severely overlapping X-ray images. As shown in Fig. 1, although X-ray images heavily overlap, each object is presented in a translucent state and perfectly preserves its characteristic information, which makes it possible to reconstruct the prohibited item from the original images. Inspired by the approach in [17] for generating synthetic X-ray images from the FB, we reserve this process and reconstruct the FB from X-ray images. However, it is difficult to train a credible foreground without the ground truth of the foreground. To address this challenge, we design a target foreground and use recursive training to adaptively approach the real foreground. Once the target foreground is created, within the restriction of the X-ray imaging characteristics, the FB can be reconstructed from the high-level features using a decoder. Finally, an attentional module is added to the detection framework to concentrate more on the foreground. The main contributions of this study are as follows:

- We propose an FBS framework, separating prohibited items as foreground from other irrelevant items and making the detection framework focus more on foreground information.
- We design the target FB to reconstruct the pure images of prohibited items in the absence of an FB ground truth.
- We synthesis a dataset with the FB ground truth and perform experiments on a synthetic dataset and two public datasets, both of which yield higher detection accuracies than current state-of-the-art methods.

The remainder of this paper is organised as follows: Section 2 outlines the related work on X-ray prohibited item detection and threat image projection. Section 3 describes in detail our FBS method. Section 4 presents the experimental setup and results. Finally, Section 5 presents the conclusions and future work.

2. Related work

In this section, we first introduce the existing methods for detecting X-ray prohibited items and then describe some methods and applications for the synthesis of X-ray prohibited item images.

2.1. X-ray prohibited item detection

Early work on X-ray object detection is generally based on bag-of-visual-words (BOVW) [19] and sparse representations [20]. Turcsany *et al.* [21] implemented bag-of-words by using the speeded-up robust features [23] detector and descriptor within a support vector machine [24] classifier framework. Their results indicate that class-specific clustering primes the feature space and ultimately simplifies the classification. Bastan *et al.* [19] presented the results of extensive experiments using BoVW with different local feature detectors and descriptors. They found that although the direct application of BoVW on X-ray images does not perform as well as regular images, the special properties in X-ray images can significantly improve the performance. Bastan *et al.* [22] extended X-ray object detection to dual-energy levels. They proposed a multiview branch-and-bound algorithm for multiview object detection. They showed that extended features and multiple views substantially improve the classification and detection performance. Fitton *et al.* [25] applied a bow model to baggage security screening in three-dimensional scenarios. They compared the effectiveness of four descriptor types and three codebook assignment methods. Franzel *et al.* [27] proposed a standard sliding-window approach with a histogram of oriented gradient [28] features in multi-view detection. Their method addresses in-plane and out-of-plane object rotation problems, and their multiview detection approach can be applied to single-view X-ray images and multiview photographic images.

Akcay *et al.* [26] first applied deep CNNs to X-ray baggage detection. They used transfer learning to a pre-trained CNN and compared the CNN approach with the BoVW approach. Their experiments demonstrated the applicability of CNNs within X-ray baggage imagery. Following the early work of [26], Mery *et al.* [29] evaluated 10 X-ray object recognition approaches based on bag-of-words, sparse representations, deep learning, and classic pattern recognition schemes. Their study indicates that deep learning methods would lead to better results in X-ray testing with large datasets and achieved the lowest computational time in the testing stage. Furthermore, Akcay *et al.* [30] explored the applicability of multiple CNN-driven detection paradigms, such as sliding window-based CNNs, Fast RCNN [31], region-based fully convolutional networks [32], and YOLOv2 [8]. They illustrated that the comparative performance of CNN-driven detection methods and object localisation strategies is better than classification techniques in cluttered X-ray security images. Focusing on complex yet meaningless contexts and class imbalance problems of existing datasets, Miao *et al.* [13] provided a dataset with heavily unbalanced positive and negative samples and they presented a CHR model, which integrates multi-level visual cues and achieves class balance through a class-balanced loss function. Their CHR method has a significant detection advantage on datasets with few positive training samples. Wei *et al.* [15] proposed DOAM to address the problem of X-ray images being heavily obscured. Their method emphasises the edge information and material information of the item and can easily be applied to existing neural network models. Hassan *et al.* [33] presented a deep cascaded multiscale structure tensor framework that extracts object information by generating a series of coherent tensors from the candidate X-ray scan. Their method can automatically extract and recognise normal and suspicious items irrespective of their position and orientation from multivendor X-ray scans.

2.2. Threat image projection

The detection performance of human screeners is heavily dependent on the experience and knowledge acquired by a large number of prohibited items [34]. This is also applicable to computer-vision object-detection algorithms. Large datasets can significantly improve the performance of the algorithm and prevent overfitting. Nevertheless, prohibited items merely account for a small portion of actual X-ray images. It takes a lot of time and money to acquire new X-ray images, so a threat image projection (TIP) method has been proposed. Rogers *et al.* [35] proposed a framework for TIP in cargo transmission X-ray imagery. The method exploits the approximate multiplicative nature of X-ray images to extract a library of threat terms and add real variation to a TIP image. In [13], Mery *et al.* used the addition of logarithmic images to replace the multiplication of the FB, which is typically used in X-ray imaging. This method allows them to use linear strategies to superimpose images of threat objects onto X-ray images. Bhowmik *et al.* [36] developed a synthetically composited data augmentation approach using TIP and investigated the differences in the detection performance of the CNN framework in detecting prohibited items under real and synthetic X-ray training images. Their evaluation demonstrates that it is promising to use synthetic images to diversify the X-ray security training imagery for automated detection algorithm training.

3. Proposed method

Inspired by the approach in [17] of composing X-ray images from the FB, we reverse the process and propose FBS to detect prohibited items in severely overlapping X-ray images. We separate the prohibited items as the foreground from other irrelevant information and use the foreground information that contains only the prohibited items to reduce the interference of other irrelevant information on detection. The architecture of the FBS is illustrated in Fig. 2. We introduce our FBS from the following aspects: 1) FBS process, 2) X-ray image modelling,

3) reconstruction of the FB, 4) the use of the attention model to make the detection framework focus on the foreground.

3.1. Motivation and formulation

Owing to the unique imaging characteristics of X-rays, objects in images overlap one another in a translucent state, and objects to be detected are often superimposed with other translucent unrelated objects. Other irrelevant objects covered on prohibited objects greatly interfere with detection. A natural solution is to separate irrelevant objects from prohibited objects. Accordingly, we assume that the prohibited items are the foreground and the other irrelevant objects are the background.

Suppose that n original input images can be represented by $O = \{o_1, o_2, \ldots, o_n\}$. After O passes through the backbone CSPDarknet53 [38] and the neck that consists of the spatial pyramid pooling [38] and path aggregation network [38], we obtain three different sizes of high-level features $E_i = \{{\varepsilon_1}^i, {\varepsilon_2}^i, \ldots, {\varepsilon_n}^i\}$ (i =1,2,3), where ${\varepsilon_j}^i$ represents the ith high-level features of the jth input image. Because the original images are complex and heavily obscured, the high-level features contain several extraneous information, which makes the detection process challenging.

Therefore, we use the foreground to refine the high-level features and exclude extraneous information. The details of reconstructing the FB and refining the high-level features are explained in Subsections 3.3 and 3.4.

The categories and locations of prohibited items are detected using the same method as YOLOv4 [38], convolving the refined high-level features to obtain heads $H = \{{h_1}^i, {h_2}^i, \ldots, {h_n}^i\}$, where ${h_j}^i$

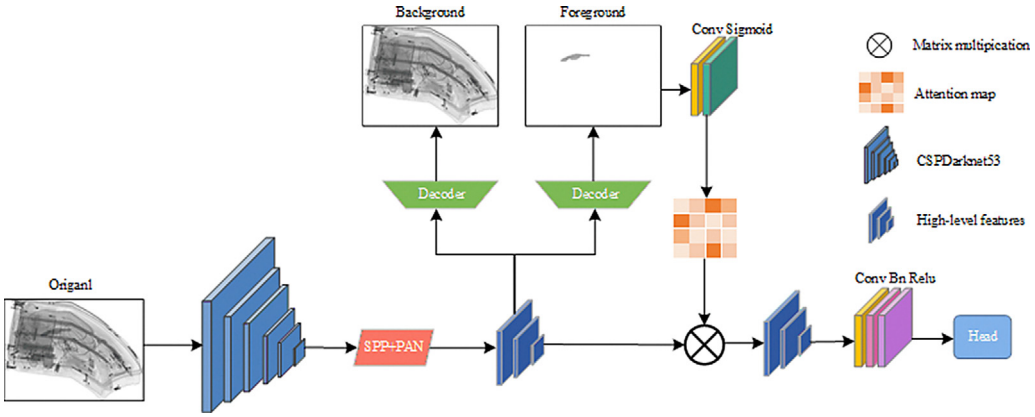

Fig. 2. Overall architecture of the proposed FBS. The high-level features are obtained through the backbone network consisting CSPDarknet53, spatial pyramid pooling, and path aggregation network, and the FB are restored from the high-level features using two decoders. The foreground attention map is generated from the foreground and multiplied with the high-level features to allow the detection framework to focus on foreground information.

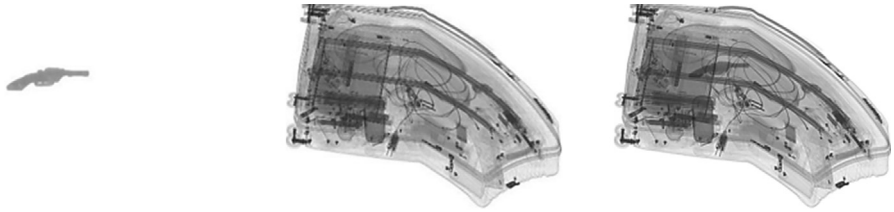

Fig. 3. X-ray image of knife superimposed onto an X-ray image of a baggage: (a) I_f: Foreground (prohibited item), (b) I_b: Background (baggage), (c) I_o: Original (baggage with the prohibited item).

represents the head of the ith high-level features of the jth input image. Every grid cell in h_j^i contains location information t_x, t_y, t_w, and t_h; confidence information C_o; and classification information C_l.

3.2. X-ray image modelling

The energy flux density loss of an X-ray penetrating an object is related to its energy, thickness, and material. When an X-ray passes through objects with uniform density and thickness, the relationship between the attenuation of the energy-flux density and matter can be expressed as

$$\varphi(d) = \varphi_0 e^{-\mu d}, \tag{1}$$

where φ is the energy flux density after passing through matter, φ_0 is the incident energy flux density, μ is the absorption coefficient and d is the thickness of the irradiated matter.

Following the intensity distribution of X-rays through matter [18], an X-ray image can be represented as

$$I = A \cdot \varphi + B \tag{2}$$

Then, the foreground of the X-ray image, e.g. a gun, and the background of the X-ray image, e.g. a baggage, as shown in Fig. 3(a) and (b), can be represented as

$$I_f = A \cdot \varphi_f + B \quad I_b = A \cdot \varphi_b + B, \tag{3}$$

where

$$\varphi_f = \varphi_0 e^{-\mu_f d_f} \quad \varphi_b = \varphi_0 e^{-\mu_b d_b}, \tag{4}$$

where μ_f and μ_b are the absorption coefficients of the foreground and background, respectively.

The total X-ray image can be modelled as

$$\varphi_o = \varphi_0 e^{-\mu_f d_f} e^{-\mu_b d_b}, \tag{5}$$

$$I_o = A \cdot \varphi_o + B = C e^{-\mu_f d_f} e^{-\mu_b d_b} + B, \tag{6}$$

where $C = A \bullet \varphi_0$. For further simplification, we obtain

$$\frac{I_o - B}{C} = \frac{I_f - B}{C} \cdot \frac{I_b - B}{C} \tag{7}$$

The normalised X-ray images can be represented as $J_o = (I_t - B) / C$. Thus, we obtain the normalised original, foreground, and background images as

$$J_o = J_f \cdot J_b \tag{8}$$

Using (8), the total image is computed as

$$I_o = C \cdot J_f \cdot J_b + B \tag{9}$$

Fig. 3. (c) shows the outcome of the TIP method.

3.3. FBS

The availability of an FB ground truth makes a big difference in the design of the algorithm for reconstructing the FB, and the FB ground truth is usually difficult to obtain in real X-ray images. In this section, we discuss cases with and without the FB ground truth.

Separation with the FB ground truth

The FB can be reconstructed from the high-level features using two decoders when the FB ground truth is available. The high-level features are upsampled five times, and each upsampled feature is combined with the corresponding feature in CSPDarknet53 and then convoluted to the same number of channels as that of the original features. The FB loss $Loss_{fb}$ can be calculated as follows:

$$Loss_{fb} = |F - F_g| + |B - B_g|, \tag{10}$$

where F and B are the reconstructed FB, and F_g and B_g are the ground truths of the FB, respectively.

Fig. 4. Target foreground and background obtained in different ways: (a) target foreground obtained by interception, (b) target background obtained by interception, (c) target foreground obtained by prediction, (d) target background obtained by prediction.

Separation without the FB ground truth

In practical X-ray detections, X-ray images usually contain only bounding box annotations instead of an FB ground truth. Relying only on (8) J_o=J_f • J_b as the restriction condition to obtain the foreground is far from sufficient. Thus, it is imperative to design a reasonable target foreground.

An intuitive solution is to use the part inside the bounding box as the target foreground and the part outside as the target background. This method is called interception. Fig. 4(a) and (b) show examples of the setup. However, this setting is rather primitive, and predicted foreground in the reconstruction is clearly better than rigidly intercepting the bounding box of the original image.

Therefore, we use recursive training to obtain the foreground when there are only bounding box annotations. This method is called prediction. In each iteration of training, we take the following steps to set the target foreground: (1) Set all the pixels outside the bounding box to 1. (2) Select 25% of the bounding box as the comparison interval for the pixels in the bounding box. If the L_1 loss of the predicted foreground of the previous iteration and the original image in the comparison interval is less than the threshold, then select the corresponding pixel in the bounding box of the predicted foreground of the previous iteration as the target foreground, as shown in Fig. 4(c) and (d). Otherwise, select the corresponding pixel in the bounding box of the original image, as shown in Fig. 4(a) and (b). The operations can be formulated as follows:

$$D = \left|F_{io} - F_{ip}\right|, \tag{11}$$

$$F_t, B_t = \begin{cases} F_o, B_o D \geq threhold \\ F_p, B_p D < threhold \end{cases}, \tag{12}$$

where F_{io} and F_{ip} represent 25% of the original image in the bounding box and 25% of the previous iteration's predicted foreground in the bounding box, respectively; D represents the difference between F_{io} and F_{ip}; F_o and B_o represent the target foreground and target background intercepted from the original image with the bounding box, respectively; F_p represents the target foreground intercepted from the previous iteration's predicted foreground with the bounding box; B_p represents the target background obtained from the previous iteration's predicted background.

The imbalance between the size inside the bounding box and the area outside the bounding box will allow the foreground reconstruction to focus more on the entire image and ignore the information in the bounding box. Therefore, we add an additional weight λ to balance the information inside and outside the bounding box. The foreground background loss $Loss_{fb}$ can be formulated as follows:

$$Loss_{fb} = |F - F_t| + |O - O_g| + \lambda\left(|F_i - F_{ti}| + \left|O_i - O_{gi}\right|\right), \tag{13}$$

$$O = F \cdot B, \tag{14}$$

where F_t is the target foreground; O_g is the ground truth of the original image, i.e. the original image; and F_i, F_{ti}, O_i, and O_{gi} denote the parts of the corresponding image in the bounding box.

In such a setting, the target foreground can adaptively approach the real foreground while ensuring that the target foreground does not differ significantly from the original image.

3.4. Attention module

We use the reconstructed foreground to refine the high-level features. Because it is redundant to train the foreground back into the network to obtain the high-level features, we use an attention module to make the high-level features focus on the foreground and eliminate irrelevant information. First, average pooling is performed on the foreground to change the number of channels to 1. Then, three features of the same size as the advanced feature, except that the number of channels is 1, are obtained by five convolutions with a kernel size of 1 × 1. The attention map is obtained using a sigmoid activation function.

Finally, the attention map is multiplied with the original high-level features.

From [38], we can compute the loss function $Loss_{lc}$ for the category and location from the refined high-level features. Then the total loss function can be expressed as

$$Loss = Loss_{lc} + Loss_{fb} \tag{15}$$

4. Experiments

In this section, we describe our extensive experiments on the proposed FBS. The experimental setup, dataset, evaluation criteria, and parameter settings are presented in detail. We then evaluate our method on three datasets and analyse the results from six aspects, namely, quantitative results, foreground visualisation, comparison with different detection approaches, ablation study, parametric analysis, and evaluation of partial images with the FB ground truth.

4.1. Experimental setup

Dataset

OPIXray. The OPIXray dataset contains a total of 8885 X-ray images of under five categories of cutters: folding knife, straight knife, scissor, utility knife, and multi-tool knife. The OPIXray dataset is especially designed for occluded prohibited item detection in security inspections. Most of the images contain only one prohibited item. Examples of OPIXray are shown in Fig. 1(b)

SIXray10. The security inspection X-ray (SIXray) [13] dataset comprises 1059231 X-ray images, 8929 of which are manually annotated for six different classes: gun, knife, wrench, pliers, scissors, and hammer. The dataset consists of objects with a wide variety of scales, viewpoints, and overlapping. We use a subset of the SIXray dataset, SIXray10, which contains 8929 positive samples and 89290 negative images. Most of the images contain multiple prohibited items. Examples of SIXray are shown in Fig. 1(c).

Synthetic dataset. Finally, we present our synthetic dataset GRIMA X-ray threat image projection dataset (GDTIPXray dataset). All original images are taken from the GRIMA X-ray (GDXray)

Table 1
Category distribution of the GDTIPXray dataset.

GDTIPXray	Categories				Total
	Gun	Shuriken	Blade	Knife	
Training	2000	2000	2000	2000	8000
Testing	500	500	500	500	2000
Total	2500	2500	2500	2500	10000

[37] and SIXray datasets [13]. GDXray is a simple dataset that contains only prohibited items. Therefore, we use the prohibited item images in GDXray as the foreground, and the messy baggage images without prohibited items in SIXray as the background to synthesise a complex dataset with the FB ground truth. Because the images selected from the SIXray dataset contain complex and tightly packed items, randomly placed prohibited items can easily overlap with other objects, making our dataset heavily overlapped.

To make the image more realistic and diverse, we randomly resize and rotate the FB images, adjust the contrast and brightness, and finally use the TIP method to generate the dataset. The synthesised GDTIPXray dataset consists of four categories: gun, shuriken, blade, and knife. The dataset contains 10000 images and is divided into a training set and testing set with a ratio of 4:1. The details of GDTIPXray are listed in Table 1, and some of its examples are shown in Fig. 1(a).

The details of the three datasets are summarised in Table 2.

Evaluation Criterion

The performances of the FBS and the prior work are evaluated by comparing the mean average precision (mAP), accuracy (ACC), macro-precision, macro-recall, and macro-F1-score (macro-F1). MAP is a metric evaluated by the area under the precision and recall curve. ACC describes the performance of the algorithm for correctly identifying the class of objects. Macro-precision, macro-recall, and macro-F1 are the extensions of precision, recall, and F1-score in the multiclass classification case. Precision is defined as TP / (TP + FP), and recall is defined as TP / (TP+ FN), where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative, respectively.

Parameter Setting

In all the following experiments, the data are first enhanced by Mosaic [38]. We use a pre-trained backbone, which is trained using VOC2007 [38]. We freeze the backbone and train it for 15 epochs and then unfreeze the backbone for 35 generations. All the models are optimised using the Adam optimiser. The initial learning rates of the frozen and unfrozen parts are 0.001 and 0.0001, respectively, and decay to 95% of the previous value in each epoch. When calculating the mAP, the threshold of the intersection over union is set to 0.5.

4.2. Experimental results

Quantitative Results

We compare our approach with other state-of-the-art methods on the GDTIPXray dataset and two publicly available datasets. To accurately compare the performance of the different methods, all algorithms choose the same backbone as YOLOv4.

1) Results on the GDTIPXray dataset: The results of different algorithms on the GDTIPXray dataset are reported in Table 3, which shows that our method outperforms the existing methods. In particular, compared with YOLOv4 [38], DOAM [15], CHR [13], and RGBS [16], our method provides an absolute gain of 3.14%, 1.87%, 1.93%, and 6.73% in terms of mAP. Among the five evaluation metrics, our FBS algorithm achieves the best results in the mAP, ACC, and macro-recall metrics. Moreover, we conduct FB (with FB) experiments when the FB ground truth is available, and our method achieves the best results in all five metrics. In this case, the reconstructed foreground is closest to the actual foreground and excludes all irrelevant information for detection, and the detection performance is the best. Furthermore, we compare the detection performances of each category. As observed in Fig. 5, the curve of our method covers the other curves in every category. These evaluations demonstrate that our FBS is effective in locating prohibited items under heavily overlapping synthetic datasets.
2) Results on the OPIXray dataset: Table 4 shows the performance of our method on the OPIXray dataset. Our FBS method outperforms YOLOv4 by 2.82%, DOAM by 2.49%, CHR by 3.12%, and RGBS by 2.72% in terms of mAP. Our algorithm also has a remarkable advantage in terms of ACC and macro-recall. In addition, on three heavily overlapping categories, i.e. straight knife, utility knife, and multi-tool knife, our method is significantly better than all the other four existing methods. The results show that our FBS is also effective in colour datasets, especially in severely overlapping categories.

1) Results on the SIXray dataset: As shown in Table 5, our method exceeds YOLOv4, DOAM, CHR, and RGBS by 1.12%, 0.47%, 0.37%, and 1.03% in terms of mAP, respectively. Our algorithm achieves optimal results for the ACC and macro-recall metrics. However, the improvement of our method on this dataset is not as prominent as that on the GDTIPXray and OPIXray datasets. This could be attributed to the fact that most images in the SIXray dataset contain multiple prohibited items, as shown in Fig. 1(c). Although the FBS method separates prohibited items from other items, the prohibited items still overlap with one another, which limits the improvement of our algorithm.

Foreground Visualisation Analysis

In this section, we visualise the foreground reconstructed from FBS to observe the effects of FBS. As shown in Fig. 6, the FBS algorithm can perfectly reconstruct the FB with the FB ground truth. FBS indeed obtains the best detection performance with the FB ground truth, which proves that the closer the reconstructed foreground is to the actual foreground, the more irrelevant information is eliminated and the more the detection performance improves.

In the samples with only bounding box annotations available, the reconstructed foregrounds are shown in Figs. 6–8. Because there are only bounding box annotations, it is difficult to achieve an accurate pixel-level reconstruction using recursive training alone. The predicted foreground can precisely locate the range of the prohibited items and reconstruct the general appearance. Although there is still a small amount of irrelevant information at the edges, the majority of the irrelevant information is excluded and the performance is significantly improved. The prohibited items in all three datasets are well separated. However, in the SIXray dataset, most images contain multiple overlapping prohibited items, which causes prohibited items in the reconstructed foreground to remain overlapped (as can be seen in the third column of Fig. 8). Thus, the FBS algorithm has a limited improvement on the SIXray dataset, which is consistent with the experimental results shown in Table 5. These visualisation results further demonstrate the effectiveness of FBS in detecting prohibited items in X-ray scanned images. Fig. 7.

Comparison with Different Detection Approaches

To further evaluate the effectiveness of FBS and to verify that FBS can be widely applied to existing detection networks, we implement FBS in the existing object detection approaches, i.e. YOLOv4 [38], SSD [11], and FCOS [39]. The results are listed in Table 6.

Table 2
Three datasets focused on different prohibited items.

Dataset	Total Samples	Training Samples	Testing Samples	Classes
GDTIPXray	10000	8000	2000	Gun, shuriken, blade, knife
OPIXray	8885	7109	1776	Folding, straight, scissor, utility, multi-tool
SIXray	98219	78575	19644	Gun, knife, wrench, plier, scissor, hammer

Table 3
Comparisons between FBS and existing algorithms on the GDTIPXray dataset

Method	AP				mAP	ACC	Macro-Precision	Macro-Recall	Macro-F1
	Blade	Gun	Knife	Shuriken					
YOLOv4	91.77	93.19	85.26	95.71	91.48	91.96	81.58	92.43	86.49
DOAM	88.79	90.34	77.37	95.06	87.89	88.92	62.93	89.82	73.66
CHR	93.95	93.52	86.17	97.86	92.87	93.23	74.19	93.76	82.69
RGBS	92.82	93.09	87.93	96.92	92.69	93.05	81.47	93.43	86.93
FBS	95.76	94.72	89.85	98.15	94.62	94.91	79.05	95.28	86.21
FBS (with FB)	**96.35**	**95.52**	**91.94**	**98.38**	**95.55**	**95.50**	**90.96**	**95.81**	**93.31**

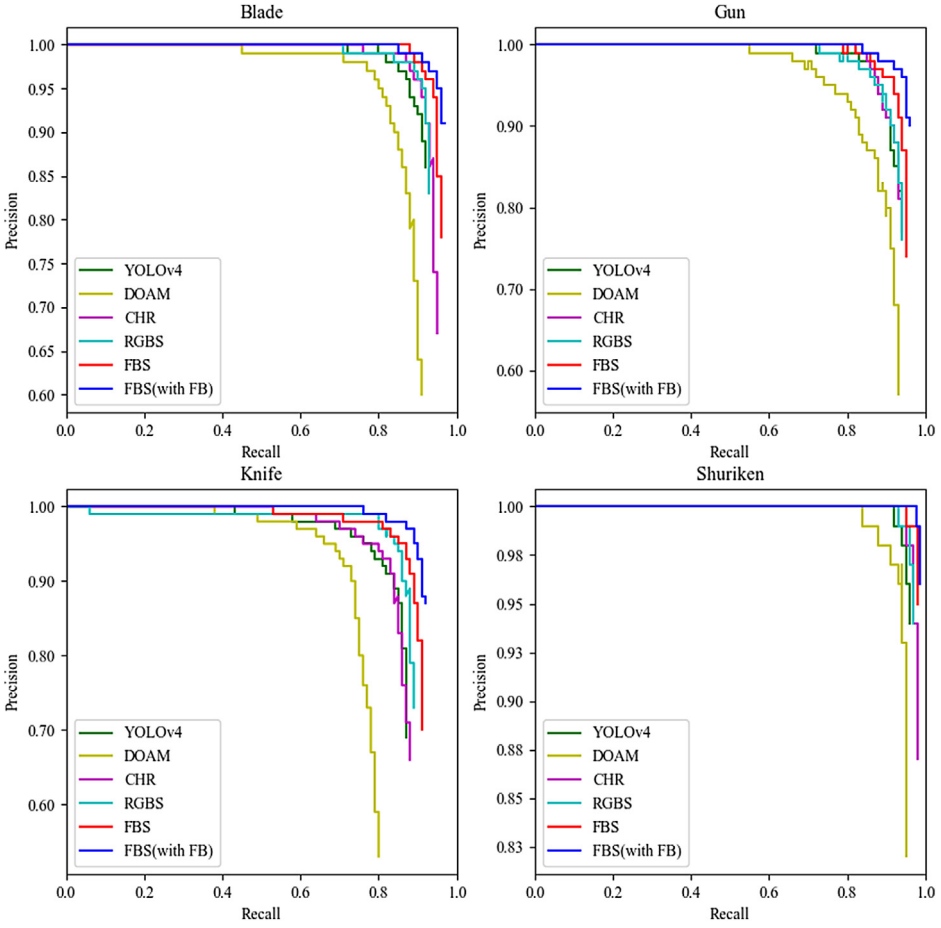

Fig. 5. Precision–recall curves obtained by different algorithms on the GDTIPXray dataset.

Table 4
Comparisons between FBS and existing algorithms on the OPIXray dataset.

Method	AP					mAP	ACC	Macro-Precision	Macro-Recall	Macro-F1
	Folding Knife	Straight Knife	Scissor	Utility Knife	Multi-tool Knife					
YOLOv4	84.57	54.91	**95.85**	75.92	83.37	78.92	86.37	69.81	84.68	76.15
DOAM	86.91	52.89	95.30	**78.34**	**84.51**	79.59	86.54	**72.62**	84.71	**78.07**
CHR	**87.94**	84.53	95.23	50.99	74.47	78.63	85.97	71.13	83.80	76.72
RGBS	85.99	84.89	95.16	54.05	75.04	79.03	86.31	68.28	84.55	75.27
FBS	86.38	**88.29**	95.45	57.99	80.62	**81.75**	**87.61**	61.71	**85.83**	67.75

Table 5
Comparisons between FBS and existing algorithms on the SIXray10 dataset.

Method	AP					mAP	ACC	Macro-Precision	Macro-Recall	Macro-F1
	Gun	Knife	Wrench	Pliers	Scissors					
YOLOv4	79.56	59.78	63.39	**75.67**	**72.17**	70.11	76.28	65.67	74.46	69.10
DOAM	**81.37**	**64.25**	73.26	70.17	61.98	70.21	**77.64**	53.02	**76.33**	61.35
CHR	79.22	63.77	73.77	71.55	65.55	70.77	76.14	67.56	74.65	70.30
RGBS	79.69	62.60	**75.09**	70.27	66.70	70.87	76.00	70.30	74.05	71.66
FBS	79.72	64.14	74.96	71.19	66.17	**71.24**	76.19	**70.37**	74.55	**71.99**

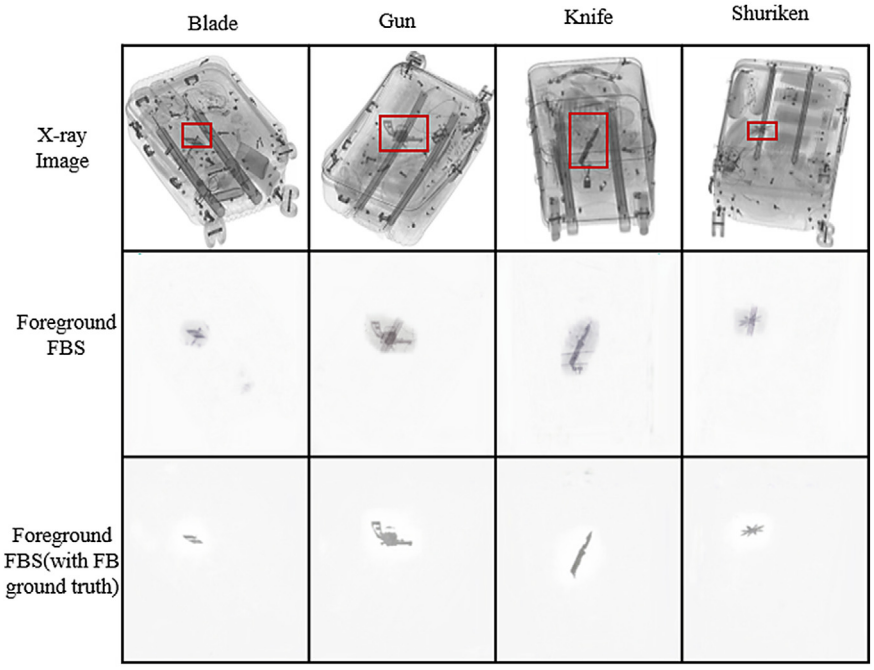

Fig. 6. Reconstruct foreground visualisation on the GDTIPXray dataset.

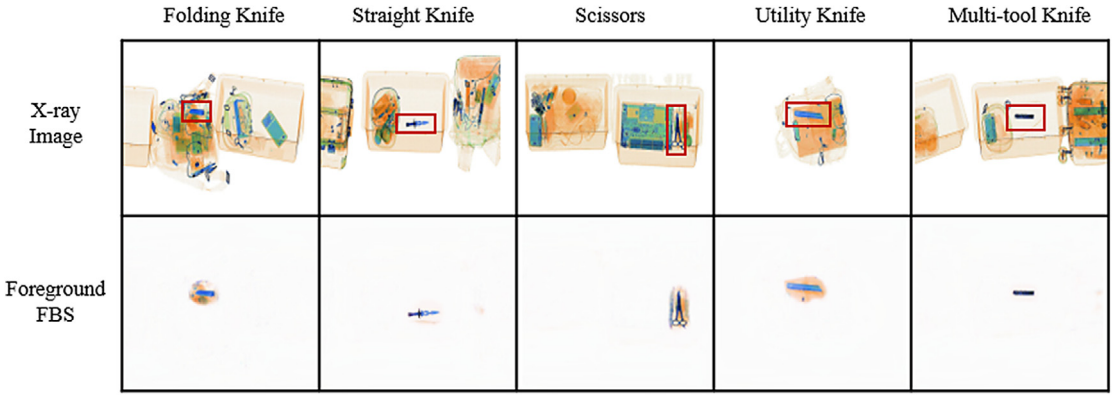

Fig. 7. Reconstructed foreground visualisation on the OPIXray dataset.

Table 6
Comparisons between the FBS-integrated network and three object detection approaches.

Method	AP				mAP	ACC	Macro-Precision	Macro-Recall	Macro-F1
	Blade	Gun	Knife	Shuriken					
YOLOv4	91.77	93.19	85.26	95.71	91.48	91.96	81.58	92.43	86.49
YOLOv4+FBS	95.76	94.72	89.85	98.15	94.62	94.91	79.05	95.28	86.21
SSD	86.22	89.22	86.26	93.96	88.91	89.29	93.64	89.46	91.45
SSD+FBS	89.73	90.66	85.78	96.98	90.79	90.88	94.10	91.26	92.65
FCOS	90.08.	91.56	84.52	96.88	90.76	88.74	87.16	89.19	88.11
FCOS+FBS	94.65	96.77	92.41	98.05	95.47	93.23	91.35	93.62	92.45

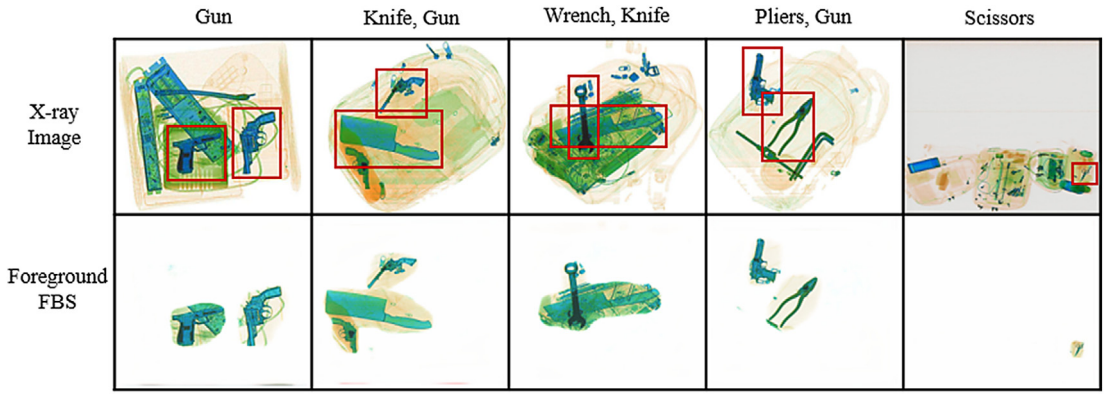

Fig. 8. Reconstructed foreground visualisation on the SIXray dataset.

Table 7
Effects of different target foreground designs on different FBS-integrated networks.

Method	AP				mAP	ACC	Macro-Precision	Macro-Recall	Macro-F1
	Blade	Gun	Knife	Shuriken					
YOLOv4+FBS+ Interception	95.46	93.28	89.2	97.56	93.88	94.23	83.08	94.63	88.3
YOLOv4+FBS+ Prediction	**95.76**	**94.72**	**89.85**	**98.15**	**94.62**	**94.91**	79.05	**95.28**	86.21
SSD+FBS+ Interception	87.61	91.75	83.72	95.76	89.71	89.56	94.28	90.12	92.11
SSD+FBS+ Prediction	**89.73**	90.66	**85.78**	**96.98**	**90.79**	**90.88**	94.10	**91.26**	**92.65**
FCOS+FBS+ Interception	93.17	82.32	82.39	97.36	91.31	87.21	88.70	88.14	88.41
FCOS+FBS+ Prediction	**94.65**	**96.77**	**92.41**	**98.05**	**95.47**	**93.23**	**91.35**	**93.62**	**92.45**

As shown in Table 6, the performance of classical detection networks in heavily overlapping X-ray scenes is significantly improved with the help of our FBS. Specifically, our method improves the mAP by 3.14%, 1.88%, and 4.71% and ACC by 2.95%, 1.59%, and 4.49% as compared to YOLOv4, SSD, and FCOS, respectively. As for the macro-precision, macro-recall, and macro-F1, our FBS-integrated network also has a more remarkable advantage as compared to SSD and FCOS. Therefore, our proposed FBS is effective and can be widely applied to existing object detection approaches.

Ablation Studies

In the FBS framework, the target foreground determines the learning direction of the reconstruction foreground. Accordingly, we investigate the impact of the designed target foreground. We perform experiments on the GDTIPXray dataset in two ways: 1) Interception method: Directly intercept the original image with the bounding box as the target FB 2) Prediction method: Use a threshold to select the original image or predicted image of the previous generation as the target FB. We compare two methods for designing the target foreground in three different FBS-integrated networks. As shown in Table 7, our designed target foreground, i.e. prediction method, achieves better results in four metrics, i.e. mAP, ACC, macro-recall, and macro-F1 under the Yolov4 and SSD detection networks, and our prediction method achieves better results in all metrics under the FCOS detection network. These experimental results demonstrate the effectiveness of the designed target foreground in FBS. The reconstructed foreground can adaptively exclude irrelevant information in the bounding box using the previous generation's predicted foreground, whereas the target foreground that directly intercepted the original image with the bounding box still preserves irrelevant information in the bounding box.

Parametric Analysis

During the training process, the threshold setting of the target foreground determines whether to select the predicted foreground of the previous generation as the target foreground, which has a significant impact on the FBS. Therefore, in this section, we compare the different threshold parameters. We perform the experiments with thresholds of 0.1, 0.2, 0.3, and 0.4, and the results are shown in Fig. 9. The best performance is achieved when the threshold is set to 0.1, which can reach 94.62% in terms of mAP. As the threshold value increases, the accuracy decreases. At thresholds of 0.2, 0.3, and 0.4, mAP decreases by 0.19%, 0.89%, and 1.44%, respectively. The results show that a higher threshold has a negative effect on the performance. When the threshold is set too high, the recursive training still tends to choose its prediction foreground as the target foreground even when there is a significant difference between its prediction foreground and the original image.

Table 8
mAP obtained by FBS at different proportions of the FB ground truth on the GDTIPXray dataset.

FB ground truth proportion	Category				mAP
	Blade	Gun	Knife	Shuriken	
100%	96.35%	95.52%	91.94%	98.38%	95.55%
10%	96.21%	95.23%	91.62%	98.18%	95.31%

Evaluation of Partial Images with the FB Ground Truth

In practice, it is worthwhile to label a small number of images with the FB ground truth to obtain the same performance as that with all FB ground truths. Therefore, we investigate the impact of adding a small number of images with the FB ground truth to the dataset. We conduct the experiment on the GDTIPXray dataset, setting up 10% of the samples with the FB ground truth, and the other 90% with only bounding box labels. As shown in Table 8, the mAP of FBS is only 0.24% lower at 10% FB ground truth than that at 100% FB ground truth, which indicates that our algorithm requires only a small number of FB ground truths to perform as well as it does with all FB ground truths.

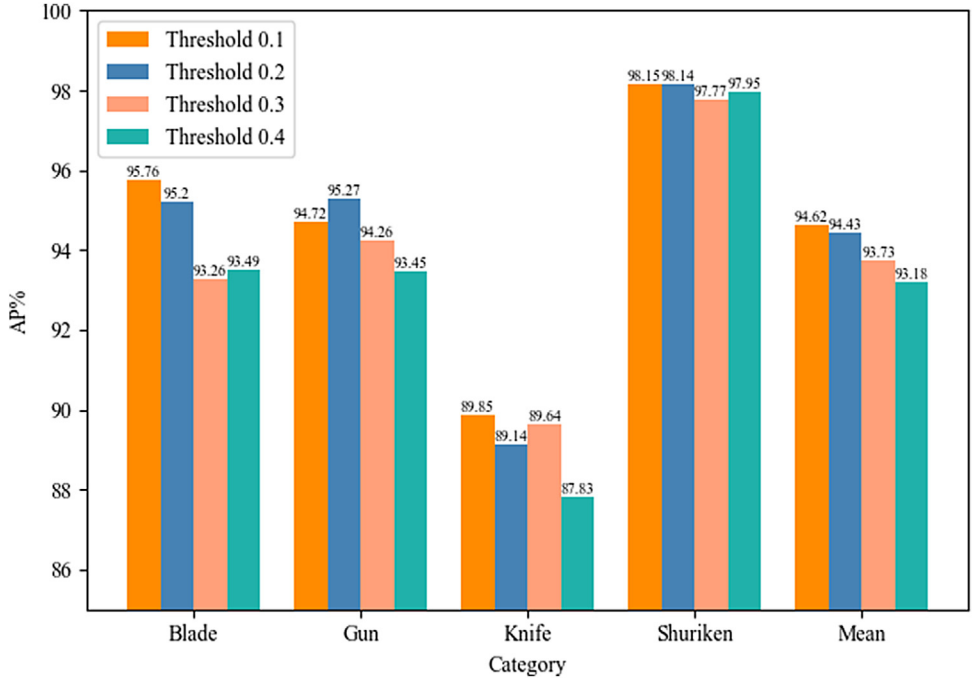

Fig. 9. Effect of different selection thresholds on the performance.

5. Conclusion

In this paper, we propose a novel method that automatically detects prohibited items in X-ray images. The proposed framework can separate prohibited items as the foreground from other irrelevant items with only the bounding box labels available and accurately identifies prohibited items of heavily obscured and overlapping X-ray images. We rigorously test the proposed algorithm on our synthetic dataset and two publicly available datasets and compare it with state-of-the-art methods. The results show that our proposed algorithm achieves the best results. Furthermore, we demonstrate the effectiveness of the target foreground designed in the proposed framework. In contrast to the rigid interception of the target foreground with the original image, the method of designing the target foreground and training it recursively to approach the real foreground significantly improves the detection performance. In addition, our method can be easily embedded into existing object detection algorithms and can improve the detection performance in X-ray scenes. Moreover, our algorithm requires only a small number of FB ground truths to achieve the same detection performance as that of all FB ground truths. In future work, to improve the performance of FBS, we will investigate semantic segmentation to separate the prohibited items in the foreground that remain overlapped. We will also extend our method to weakly supervised scenes by separating the FB without bounding box labels.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Key Project of Science and Technology Innovation 2030 supported by the Ministry of Science and Technology of China under Grant 2018AAA0101302, and in part by the General Program of National Natural Science Foundation of China (NSFC) under Grant 61773300.

References

[1] D. Mery, in: "Computer vision for X-Ray testing,", 10, Springer International Publishing, Switzerland, 2015, pp. 973–978.
[2] P. Wu, J. Liu, F. Shen, Y. Shi, Y. Sun, F. Shao, Z. Wu, Z. Yang, Not only look, but also listen: learning multimodal violence detection under weak supervision, in: European Conference on Computer Vision, 2020, pp. 322–339.
[3] P. Wu, J. Liu, M. Li, Y. Sun, F. Shen, Fast sparse coding networks for anomaly detection in videos, Pattern Recognit. 107 (2020) 107515.
[4] P. Wu, J. Liu, F. Shen, A deep one-class neural network for anomalous event detection in complex scenes, IEEE Trans. Neural Netw. Learn. Syst. 31 (7) (July 2020) 2609–2622.
[5] J. Li, G. Zhao, Y. Tao, P. Zhai, H. Chen, H. He, T. Cai, Multi-task contrastive learning for automatic CT and X-ray diagnosis of COVID-19, Pattern Recognit. 114 (2021) 107848.
[6] Z. Wang, Y. Xiao, Y. Li, J. Zhang, F. Lu, M. Hou, X. Liu, Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays, Pattern Recognit. 110 (2020) 107613.
[7] S. Akcay, T.P. Breckon, An evaluation of region based object detection strategies within X-ray baggage security imagery, in: 2017 IEEE International Conference on Image Processing, 2017, pp. 1337–1341.
[8] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 39 (6) (2017) 1137–1149.
[9] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.
[10] K.J. Liang, G. Heilmann, C. Gregory, S.O. Diallo, D. Carlson, G.P. Spell, J.B. Sigman, K. Roe, L. Carin, Automatic threat recognition of prohibited items at aviation checkpoint with x-ray imaging: a deep learning approach, *Anomal. Detect. Imag. X-Rays* III 10632 (2018) 1063203.
[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, 2016, pp. 21–37.
[12] Y.F.A. Gaus, N. Bhowmik, S. Akcay, T. Breckon, Evaluating the transferability and adversarial discrimination of convolutional neural networks for threat object detection and classification within X-ray security imagery, in: 2019 18th IEEE International Conference on Machine Learning and Applications, 2019, pp. 420–425.
[13] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, Q. Ye, Sixray: a large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2119–2128.
[14] O. Li, P.L. Shui, Color edge detection by learning classification network with anisotropic directional derivative matrices, Pattern Recognit. (2021) 108004.

[15] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, X. Liu, Occluded prohibited items detection: an X-ray security inspection benchmark and de-occlusion attention module, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 138–146.
[16] J. Liu, X. Leng, Y. Liu, Deep convolutional neural network based object detector for X-ray baggage security imagery, in: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence, 2019, pp. 1757–1761.
[17] D. Mery, A.K. Katsaggelos, A logarithmic x-ray imaging model for baggage inspection: Simulation and object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 57–65.
[18] H.E. Martz, C.M. Logan, D.J. Schneberk, P.J. Shull, X-ray Imaging: Fundamentals, Industrial Techniques and Applications, CRC Press, 2016.
[19] M. Baştan, M.R. Yousefi, T.M. Breuel, Visual words on baggage X-ray images, in: International Conference on Computer Analysis of Images and Patterns, 2011, pp. 360–368.
[20] D. Mery, E. Svec, M. Arias, Object recognition in baggage inspection using adaptive sparse representations of X-ray images, in: *Pacific-Rim Symposium on Image and Video Technology* 2015: *Image and Video Technology*, 2015, pp. 709–720.
[21] D. Turcsany, A. Mouton, T.P. Breckon, Improving feature-based object recognition for X-ray baggage security screening using primed visualwords, in: 2013 IEEE International Conference on Industrial Technology, 2013, pp. 1140–1145.
[22] M. Baştan, Multi-view object detection in dual-energy X-ray images, Mach. Vision Appl. 26 (7-8) (2015) 1045–1060.
[23] H. Bay, A. Ess, T. Tuytelaars, V. Gool, "Speeded-up robust features (SURF). Computer vision and image understanding (CVIU), Proc of the 9th European Conference on Computer Vision, Springer, Austria, 2006.
[24] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Appl. 13 (4) (1998) 18–28.
[25] G. Flitton, A. Mouton, T.P. Breckon, Object classification in 3D baggage security computed tomography imagery using visual codebooks, Pattern Recognit. 48 (8) (2015) 2489–2499.
[26] S. Akçay, M.E. Kundegorski, M. Devereux, T.P. Breckon, Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery, in: 2016 IEEE International Conference on Image Processing, 2016, pp. 1057–1061.
[27] T. Franzel, U. Schmidt, S. Roth, Object detection in multi-view X-ray images, in: Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium, 2012, pp. 144–154.
[28] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, 2005, pp. 886–893.
[29] D. Mery, E. Svec, M. Arias, V. Riffo, J.M. Saavedra, S. Banerjee, Modern computer vision techniques for x-ray testing in baggage inspection, IEEE Trans. Syst. Man Cybernetic. 47 (4) (2016) 682–692.
[30] S. Akcay, M.E. Kundegorski, C.G. Willcocks, T.P. Breckon, Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery, IEEE Trans. Inf. Forensics Secur. 13 (9) (2018) 2203–2215.
[31] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[32] J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, in: Advances in Neural Information Processing Systems, 2016, pp. 379–387.
[33] T. Hassan, S. H. Khan, S. Akcay, M. Bennamoun, and N. Werghi, "Deep CMST Framework for the autonomous recognition of heavily occluded and cluttered baggage items from,multivendor security radiographs," *arXiv preprint arXiv1912.04251*, 2019.
[34] S. Michel, S.M. Koller, J.C. de Ruiter, R. Moerland, M. Hogervorst, A. Schwaninger, Computer-based training increases efficiency in X-ray image interpretation by aviation security screeners, in: 2007 41st Annual IEEE International Carnahan Conference on Security Technology, 2007, pp. 201–206.
[35] T.W. Rogers, N. Jaccard, E.D. Protonotarios, J. Ollier, E.J. Morton, L.D. Griffin, Threat Image Projection (TIP) into X-ray images of cargo containers for training humans and machines, in: 2016 IEEE International Carnahan Conference on Security Technology (ICCST), 2016, pp. 1–7.
[36] N. Bhowmik, Q. Wang, Y.F.A. Gaus, M. Szarek, T.P. Breckon, The good, the bad and the ugly: Evaluating convolutional neural networks for prohibited item detection using real and synthetically composited X-ray imagery, British Machine Vision Conference Workshop, 2019.
[37] D. Mery, V. Riffo, U. Zscherpel, G. Mondragón, I. Lillo, I. Zuccar, H. Lobel, M. Carrasco, "GDXray: The database of X-ray images for nondestructive testing, J. Nondestr. Eval. 34 (4) (2015) 42.
[38] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv2004.10934*, 2020.
[39] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9627–9636.

Fangtao Shao received the B.S. degree in Mathematics and Applied Mathematics from Xidian University, Xi'an, China, in 2019, where he is currently pursuing the Master degree with the School of Artificial Intelligence. His current research interests include anomaly detection, object detection and deep learning.

Jing Liu received the B.S. degree in computer science and technology and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2000 and 2004, respectively. In 2005, she joined Xidian University as a Lecturer, where she was promoted to a Full Professor in 2009. From April 2007 to April 2008, she was a Post-Doctoral Research Fellow with The University of Queensland, Brisbane, QLD, Australia. From July 2009 to July 2011, she was a Research Associate with The University of New South Wales, Australian Defense Force Academy, Canberra, ACT, Australia. She is currently a Full Professor with the School of Artificial Intelligence, Xidian University. Her current research interests include evolutionary computation, complex networks, fuzzy cognitive maps, multiagent systems, and data mining. Dr. Liu serves as an Associate Editor for the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. From 2017 to 2018, she was the Chair of Emerging Technologies Technical Committee (ETTC) of the IEEE Computational Intelligence Society.

Peng Wu received the B.Eng. degree from Xidian University, Xi'an, China, in 2017, where he is currently pursuing the Ph.D. degree with the School of Artificial Intelligence. His current research interests include anomaly detection, weakly supervised learning, and deep learning.

Zhiwei Yang received the B.S. degree in Communication Engineering from Zhongyuan University of Technology, Zheng'zhou, China, in 2019 and now he is currently pursuing the Master degree in Electronics and Communication Engineering from Xidian University, Xi'an, China. His current research interests include computer vision, anomaly detection and deep learning.

Zhaoyang Wu received the B.S. degree in intelligent science and technology in 2019 from Xidian University, Xi'an, China, where he is currently working toward the Academic master's degree with Guangzhou Institute of Technology, Xidian University, Xi'an 710071, China. His current research interests include deep learning, anomaly detection, data mining, and evolutionary computation.