Neural Computing and Applications (2022) 34:20285–20298
https://doi.org/10.1007/s00521-022-07578-7

ORIGINAL ARTICLE

Handling occlusion in prohibited item detection from X-ray images

Dongsheng Liu[1] · Yan Tian[1] · Zhaocheng Xu[1] · Guotang Jian[1]

Received: 19 March 2022 / Accepted: 28 June 2022 / Published online: 21 July 2022

Abstract
Prohibited item detection from X-ray images determines whether any prohibited items are present in baggage, and great progress has recently been made in this field with the development of deep learning. Nevertheless, the appearance of an occluded item interacts with the cover, which is different from occlusions encountered in conventional object detection. We design three mechanisms to handle this challenge on the assumption that the occluded part is still partially observed. First, we propose a scale interaction module in which the features in neighboring scales interact one or more times to enhance the model's perception ability. Then, we design a cross-image weakly supervised semantic analysis model utilizing the coattention mechanism to perceive similar and different targets, breaking through the information bottleneck of the isolated detection of a single image. Finally, we introduce a multitask learning module to simultaneously optimize the model at the global level and pixel level. We evaluate our approach on the publicly available security inspection X-ray (SIXray) dataset, the occluded prohibited items X-ray (OPIXray) dataset, and the HIXray dataset, and the results show that our approach is competitive with other X-ray baggage inspection approaches.

Keywords Computer vision · Convolutional neural network · Information fusion · Weakly supervised learning

1 Introduction

Prohibited item detection from X-ray images can automatically search for prohibited items in passenger packages, thereby effectively suppressing terrorism and criminal incidents. Compared to other nondestructive detection methods (such as ultrasound, overfrequency imaging, and thermal imaging), the advantage of this technique lies in its excellent recognition, clarity and visualization abilities. Therefore, intelligent prohibited item detection based on X-ray images has always been a popular area of research in the multimedia field.

Recently, deep learning, especially deep convolutional neural networks [3, 7, 13, 29], has been successfully applied to prohibited item detection. However, occlusion in X-ray baggage inspection is different from that in conventional object detection. The occluded parts are totally invisible in conventional object detection scenarios, while the occluded items can still be observed in detection based on X-ray images. Examples are illustrated in Fig. 1. The appearance of an item in an X-ray image depends not only on the specific item but also on the interacting item. To solve the occlusion problem in prohibited item detection from X-ray images, several approaches [7, 29] use edge information to enhance the model's discrimination capacity. However, the gradient information introduces too much noise, which causes high uncertainty due to inference. In addition, models for extracting semantic edge features require supervised learning, but the true labels used for model learning can be obtained only through a complex and tedious labeling process. Therefore, there is an urgent need for a detection method that can combine semantic information to detect partially visible prohibited items without the cumbersome labeling process.

This work was supported in part by the National Natural Science Foundation of China under Grant 61972351 and 62111530300, in part by the Public Welfare Technology Research Project of Zhejiang Province under Grant LGF19G010002 and LGF20G010002, in part by the Science and Technology Program of Zhejiang Province (Key Research and Development Plan) under Grant 2022C01005, and in part by the Special Project for Basic Business Expenses of Zhejiang Provincial Colleges and Universities under Grant JRK22003.

✉ Yan Tian
tianyan@zjgsu.edu.cn

[1] School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China

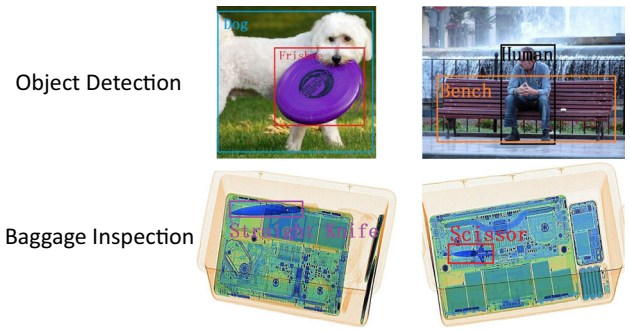


Fig. 1 Illustration of the occlusion difference between prohibited item detection from X-ray images and general object detection. Occluded pixels are unobserved in general object detection, while prohibited items from X-ray images should overlap with other items

We argue that 3 factors may help to uncover the occluded prohibited item in an X-ray image: (1) Multiscale analysis, since other visible parts of an item provide valuable knowledge for learning. (2) The characteristics of the object that distinguish it from other objects such that the object is discovered only when these characteristics are found. (3) Multitask learning that utilizes the association between tasks to assist the inference, e.g., segmentation and detection; however, since pixelwise labeling requires complicated annotation for learning, multitask learning without the tedious labeling process is particularly attractive. Thus, we must design an approach to detect partially observed prohibited items by combining these factors.

We propose a method for detecting prohibited objects based on X-ray images, as shown in Fig. 2. The scale interaction module (SIM) extracts features through the encoder by simultaneously exploring information at multiple scales. Then, the cross-image analysis module (CAM) uses the coattention mechanism to discriminate the semantics by using object images from the same or different classes, which provides weakly supervised information for localization. Finally, the multitask learning module (MLM) simultaneously learns the localization and segmentation branches, in which the segmentation branch is learned in a weakly supervised manner to alleviate the annotation effort.

The novelty of this work is exemplified by the following:

- The context information is explored by incorporating features from neighboring scales to improve the discrimination capabilities.
- Cross-image semantics are introduced to further extract high-level knowledge by using two different coattentions.
- A weakly supervised method is proposed to learn the segmentation branch in an MLM.

The experimental results on the security inspection X-ray (SIXray) dataset [13], the occluded prohibited items X-ray (OPIXray) dataset [29], and the HIXray dataset [16] show that our approach outperforms other state-of-the-art prohibited item detection approaches by margins of 1.47 and 1.45 in mean average precision (mAP).

The rest of this paper is organized as follows. Section 2 provides an overview of the recent work on prohibited item detection and general object detection. Section 3 introduces our approach. Section 4 discusses the experimental results. Conclusions are reported in Sect. 4.

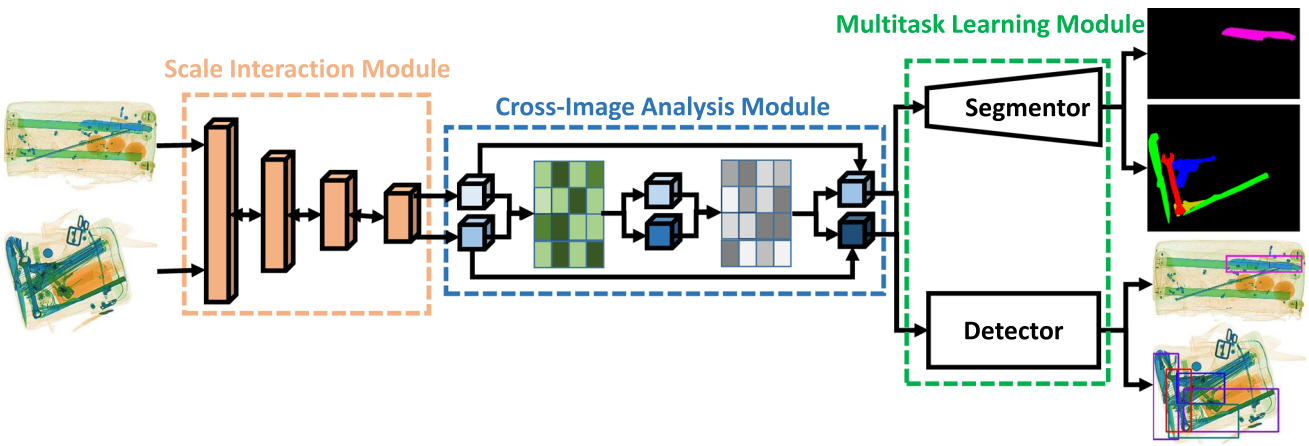


Fig. 2 Illustration of the framework. SIM extracts features through the encoder by simultaneously exploring information at multiple scales. Then, the CAM uses the coattention mechanism to discriminate the semantics by using object images from the same or different classes, which provides weakly supervised information for localization. Finally, the MLM simultaneously obtains the localization and segmentation outputs

2 Related work

In this section, we briefly review the related research on prohibited item detection and general object detection.

2.1 General object detection

Deep learning has made great achievements in the field of general object detection [19, 22]. According to the measure of whether there is a candidate anchor generation stage, these methods can be divided into the following two categories: (1) Single-stage approaches, such as the scaled you only look once version 4 (S-YOLOv4) [27], which directly regress the object location and category from all candidate locations, have advantages in efficiency. However, invalid candidates occupy a large proportion, which decreases the effectiveness of this kind of approach. (2) Two-stage approaches, such as the faster region-based convolutional neural network (Faster R-CNN) [15], first locate candidate anchors and distinguish foreground and background regions, and then the category and location of the candidate anchor is determined. These approaches have higher accuracy. Nevertheless, the initial positioning of objects requires extensive calculations, and as a result, these approaches may be slow [24].

A large number of diverse samples is important for training. Copy-Paste [5] pastes objects from one image to another image, which is a useful mechanism for data augmentation. To solve the sample imbalance problem, RetinaNet [10] employs focal loss to reduce the contribution of easy samples.

Multiscale analysis perceives the input from different perception fields [21], among which class-balanced hierarchical refinement (CHR) [13] and recursive feature pyramid and switchable atrous convolution detection (DetectoRS) [14] incorporate extra feedback connections from high-level features to improve the semantic features.

Contextual exploration has gradually become a popular research topic [17, 20], especially mining nonlocal dependent information, such as nonlocal neural networks [28], dual attention networks (DANs) [4] and ternary attention networks [25]. Recently, the transformer [26] has become a prevalent model architecture. The shifted window (swin transformer) [12] constructs a hierarchical representation to expand the applicability of transformers. The focal transformer [30] performs fine-grained self-attention only in local regions and coarse-grained attention globally.

However, research on learning the discrepancies of different semantic features and the interactions between different feature maps is very limited. In addition, due to the loss of appearance and geometric information, coupled with the limited ability to extract semantic information, the above methods are very sensitive to overlap and occlusion phenomena. Moreover, the attention mechanism used by the above methods considers only the isolated information in a single image to assign pixel weights, which makes the detection model extremely vulnerable to the bottleneck constraint of a single image and thus deviates from the overall distribution of the dataset, significantly reducing the effectiveness in scenarios where an object is highly occluded, such as in package inspection.

2.2 Prohibited item detection

For prohibited item detection from X-ray images, transfer learning is introduced to learn the differences between general detection tasks and prohibited item detection [2]. To alleviate the adverse effects of outliers, joint learning of high-dimensional image generation and spatial reasoning based on a conditional generative adversarial network [1] is studied. To alleviate the negative impact of complex scenes, CHR [13] investigates the effectiveness of sample balance and multiscale analysis in prohibited item detection.

Because package capacity is inevitably limited, the items in packages are highly occluded and overlap. In this regard, Wei et al. [29] proposed a deocclusion attention module (DOAM) based on appearance, material and color information to extract edge information for the detection process. The work employing the cascaded structure tensor (CST) [7] uses a similar idea, fusing gradients in different directions in an iterative manner. The work [9] combines CST and transfer learning to further address occlusion problems and achieves desirable results.

However, the edge cues contain too many irrelevant gradients. Therefore, they do not improve localization and classification capabilities, causing the detection model to have poor discrimination ability in cases of severe occlusion. In addition, the above methods are limited to using single images for training, which prevents the model from using other potential information in different images to form a more comprehensive understanding of prohibited and nonprohibited objects.

3 Our approach

To handle occlusion in prohibited item detection from X-ray images without introducing a cumbersome and complicated labeling process, we propose a weakly supervised learning method, as shown in Fig. 1.

3.1 Scale interaction module

Multiscale analysis provides the potential to handle ambiguity challenges because other parts of an occluded item provide valuable information for judgment [23]. However, the current approach [29] combines features of different scales only by a linear combination. This manner of combination does not take the large semantic gaps among different perception fields into consideration. Therefore, the fused features have a lower discrimination capacity because of the inconsistent semantic information.

In fact, there is a dependency between feature maps of adjacent scales (as shown in Fig. 3). Thus, we model this dependency and utilize it for baggage inspection.

We propose a method named SIM, as illustrated in Fig. 4, to improve the feature discernment of the model with interacting feature maps of adjacent scales to avoid training fluctuations caused by semantic gaps.

First, assume that the input of the module is image $\mathbf{I}$, and that the multiscale initial feature maps $\{\mathbf{f}_0^1, \mathbf{f}_0^2, \ldots, \mathbf{f}_0^C\}$ are obtained by the encoder, where C is the number of feature map scales, which is selected in the experiments. The encoder is a residual learning network composed of C residual blocks [8]. While additional scales for interaction would obtain additional global information, this would also increase the risk of overfitting. Each residual block is composed of a batch normalization layer, a rectified linear unit, and a convolutional layer with a kernel size of 3×3.

To mitigate the semantic gap, we acquire interactions among feature maps of neighboring scales; that is, low-level feature maps $\mathbf{f}_0^{i-1}$, middle-level feature maps $\mathbf{f}_0^i$, and high-level feature maps $\mathbf{f}_0^{i+1}$ are aggregated to obtain multiscale interaction maps. The scale interaction process can be added by projection as follows:

$$\mathbf{f}_{k+1}^i = \mathbf{W}_{down}^{i-1}\mathbf{f}_k^{i-1} + \mathbf{W}_k^i\mathbf{f}_k^i + \mathbf{W}_{up}^{i+1}\mathbf{f}_k^{i+1}, \quad (1)$$

where $\mathbf{W}_{up}^{i+1}$ and $\mathbf{W}_{down}^{i-1}$ represent the upsampling and downsampling operations implemented by the 3×3 deconvolutional and convolutional layers with a stride of 2, respectively. All channel numbers change to the same number as the i-th scale feature maps.

To explore nonlocal dependencies, the scale interaction process can be performed one or more times to generate multiscale interaction feature maps. We take the feature interaction maps $\mathbf{f}_1^i, \ldots, \mathbf{f}_{k-1}^i$ generated in round $1 : k-1$ as the input to the round k interaction. The total number of iterations is K.

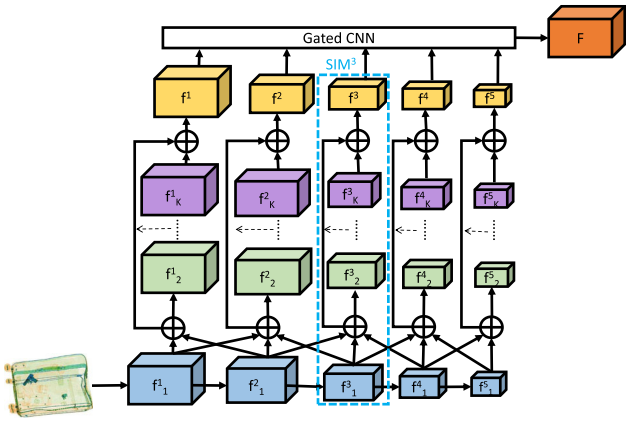


Fig. 4 SIM Illustration. For simplification, only the third SIM module has a blue dotted box

In addition, a residual learning strategy is introduced to prevent vanishing gradient issues, and pixel summation is performed on all K round feature interaction maps generated at scale i to obtain the context feature maps $\{\mathbf{f}^i\}, i = 1, 2, \ldots, C$ at each scale. Finally, the branches that pertain to different scales are fused together through a gated CNN [31] because the context feature maps in each branch contain information regarding a specific perception field.

Multiscale feature fusion via an SIM has the following advantages: (1) The feature interaction of neighboring scales enhances the feature representation, alleviates the semantic gaps between features in different perception domains, and allows the model to obtain a comprehensive understanding of the entirety and different parts of the same object. (2) It can effectively capture the appearance variations caused by severe occlusion, where the visual features are seriously insufficient. (3) The SIM can be directly used as a plug-and-play module in various applications; moreover, it is efficient and easy to train.

3.2 Cross-image analysis module

The unique attributes of an object are also an important basis for detecting it and can be used for distinguishing a particular kind of object from other objects. The key point is how to distinguish and locate the most unusual parts of prohibited items.

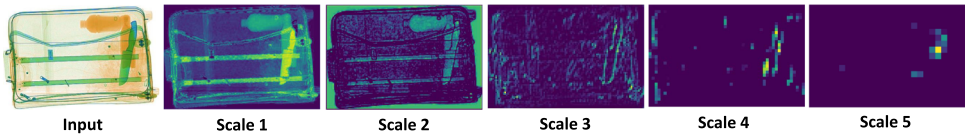


Fig. 3 Example of the relation between feature maps from neighboring scales

Inspired by the notion of identifying a new object by comparing it with a reference, such as a photo or an explanation text for a specific class, we attempt to understand the attributes and patterns of prohibited items by comparing a target image with reference images through a cross-image attention mechanism to explore the characteristics between them and weaken complex background interference. The common attention explores cross-image shared semantics, which helps the classifier to proficiently perceive the common semantic labels over the coattentive regions. Discrepancy attention focuses on unshared semantics, which enables the classifier to capably separate the semantic patterns of different objects.

Moreover, we focus on locating unique areas by using weakly supervised signals; that is, the classification task is employed to discover the unique part of prohibited items by optimizing the cross-entropy of the common class and discrepancy class in the image pair. Compared to the detection task, this weakly supervised learning method has relatively less labeling effort. Cross-image semantic relations are used as additional category-level information to guide the learning stage. Specifically, we design a CAM using common attention and discrepancy attention mechanisms to learn cross-image semantic representations for prohibited items. The details are shown in Fig. 5.

Assuming that image $\mathbf{I}_m$ is a target image and $\mathbf{I}_n$ is a reference image that is randomly selected from a reference set containing at least one kind of prohibited item with the target image, we resize these two images to a fixed size. The symbol $\mathbf{l}_n \in \{0,1\}^K$ represents a category label corresponding to $\mathbf{I}_n$ (elements corresponding to prohibited items in an image are denoted as '1', and remaining elements in the label vector are denoted as '0'), and K equals the number of prohibited item categories. The feature maps $(\mathbf{F}_m, \mathbf{F}_n) \in R^{U\times H\times W}$ obtained by the SIM are used as the CAM input, where U, H, and W are the number of channels, height, and width of the feature map, respectively. Then, the feature maps $(\mathbf{F}_m, \mathbf{F}_n)$ are processed by class-aware full convolution (CFC) to obtain the activation maps $(\mathbf{S}_m, \mathbf{S}_n) \in R^{Q\times H\times W}$, where Q is the number of channels of the activation maps. After that, the category score vectors $(\mathbf{s}_m, \mathbf{s}_n) \in R^Q$ are obtained through global average pooling. Finally, the sigmoid function is applied to normalize and obtain the cross-entropy loss function L_{ce}. The single-image classification loss of the image pair is as follows:

$$L_{single}^{m,n} = L_{ce}(\mathbf{s}_m, \mathbf{l}_m) + L_{ce}(\mathbf{s}_n, \mathbf{l}_n). \tag{2}$$

To learn common attention in an image pair, the feature maps $(\mathbf{F}_m, \mathbf{F}_n)$ are reshaped to obtain flattened feature maps $(\overline{\mathbf{F}}_m, \overline{\mathbf{F}}_n) \in R^{U\times HW}$, where HW is the number of pixels in the input feature map. The cross-image common attention similarity matrix

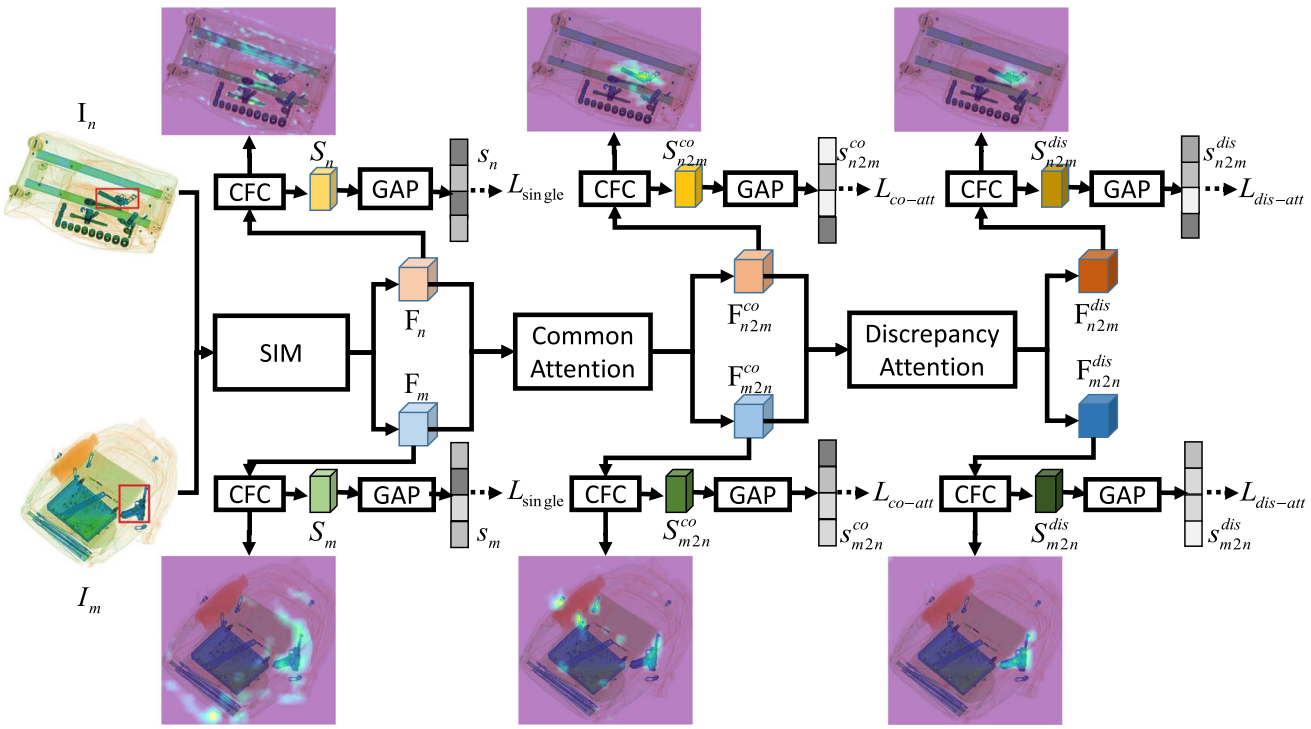


Fig. 5 Details of the cross-image semantic relation exploration. The feature maps are obtained via the SIM; then, the context features are generated by using the coattention mechanism

$$\mathbf{P}_{mn} = \overline{\mathbf{F}}_m^T \mathbf{W}_p \overline{\mathbf{F}}_n, \tag{3}$$

$$\mathbf{P}_{nm} = \overline{\mathbf{F}}_n^T \mathbf{W}_p \overline{\mathbf{F}}_m, \tag{4}$$

is used to measure the similarity between any positions of two different feature maps, where $\mathbf{W}_p \in R^{U\times U}$ is the weight matrix to be learned. $\mathbf{P}_{mn}, \mathbf{P}_{nm} \in R^{HW\times HW}$ are normalized with the softmax function to obtain a cross-image common attention map $(\mathbf{A}_m, \mathbf{A}_n) \in R^{HW\times HW}$, and then the flattened cross-image common context feature maps are obtained

$$\overline{\mathbf{F}}_{m2n}^{co} = \overline{\mathbf{F}}_m \mathbf{A}_m \in R^{U\times HW}, \tag{5}$$

$$\overline{\mathbf{F}}_{n2m}^{co} = \overline{\mathbf{F}}_n \mathbf{A}_n \in R^{U\times HW}. \tag{6}$$

We adjust the shape of $(\overline{\mathbf{F}}_{m2n}^{co}, \overline{\mathbf{F}}_{n2m}^{co})$ to obtain common context feature maps $(\mathbf{F}_{m2n}^{co}, \mathbf{F}_{n2m}^{co}) \in R^{U\times H\times W}$. The class-aware activation maps $(\mathbf{S}_{m2n}^{co}, \mathbf{S}_{n2m}^{co}) \in R^{Q\times H\times W}$ are obtained through the CFC, and the class score vectors $(\mathbf{s}_{m2n}^{co}, \mathbf{s}_{n2m}^{co}) \in R^{Q}$ are obtained through global average pooling. Finally, the cross-image common attention classification loss is calculated using the sigmoid cross-entropy

$$L_{co-att}^{m,n} = L_{ce}(\mathbf{s}_{m2n}^{co}, \mathbf{l}_m \cap \mathbf{l}_n) + L_{ce}(\mathbf{s}_{n2m}^{co}, \mathbf{l}_n \cap \mathbf{l}_m), \tag{7}$$

where $\mathbf{l}_m \cap \mathbf{l}_n$ is the common category label of image pair $(\mathbf{I}_m, \mathbf{I}_n)$.

To understand the objects well, we also learn the discrepancy attention by exploring the semantic difference between different objects in the image pair.

Assuming that the parameter matrix $\mathbf{W}_b \in R^{1\times U}$ to be learned collects common semantic knowledge (implemented by a 1×1 convolutional layer), the sigmoid activation function is $\sigma(.)$; then, the class-independent attention maps are

$$\mathbf{B}_{m2n}^{co} = \sigma(\mathbf{W}_b \mathbf{F}_{m2n}^{co}), \tag{8}$$

$$\mathbf{B}_{n2m}^{co} = \sigma(\mathbf{W}_b \mathbf{F}_{n2m}^{co}). \tag{9}$$

The discrepancy attention maps of the unshared semantic region can be obtained by

$$\mathbf{A}_{m2n}^{dis} = 1 - \mathbf{B}_{m2n}^{co}, \tag{10}$$

$$\mathbf{A}_{n2m}^{dis} = 1 - \mathbf{B}_{n2m}^{co}, \tag{11}$$

then the discrepancy context feature can be obtained by

$$\mathbf{F}_{n2m}^{dis} = \mathbf{F}_m \otimes \mathbf{A}_{n2m}^{dis}, \tag{12}$$

$$\mathbf{F}_{m2n}^{dis} = \mathbf{F}_n \otimes \mathbf{A}_{m2n}^{dis}, \tag{13}$$

where $\otimes$ is the elementwise product. Likewise, the activation maps $(\mathbf{S}_{m2n}^{dis}, \mathbf{S}_{n2m}^{dis}) \in R^{Q\times H\times W}$ can be obtained through CFC, the category score vectors $(\mathbf{s}_{m2n}^{dis}, \mathbf{s}_{n2m}^{dis}) \in R^{Q}$ can be obtained through global average pooling, and the cross-image discrepancy attention classification loss is as follows:

$$L_{dis-att}^{m,n} = L_{ce}(\mathbf{s}_{n2m}^{dis}, \mathbf{l}_m \backslash \mathbf{l}_n) + L_{ce}(\mathbf{s}_{m2n}^{dis}, \mathbf{l}_n \backslash \mathbf{l}_m), \tag{14}$$

where $\mathbf{l}_m \backslash \mathbf{l}_n$ represents the object classes that exist in image $\mathbf{I}_m$ and that do not exist in $\mathbf{I}_n$, and likewise for $\mathbf{l}_n \backslash \mathbf{l}_m$. Finally, the total training loss function of the weakly supervised learning is as follows:

$$L_{total} = \sum_{m,n} [L_{single}^{m,n} + \alpha(L_{co-att}^{m,n} + L_{dis-att}^{m,n})], \tag{15}$$

where α is the weight of the cross-image attention classification loss.

The usage of common and discrepancy attention in the CAM has the following advantages: (1) The rich contextual semantic information between images is explored by means of common/discrepancy attention to understand the unique parts of prohibited items. (2) The weakly supervised signals from the class-aware activation map reduce the tedious labeling process. (3) Due to the large number of configurations in image pairs, this approach works similarly to data augmentation to improve semantic understanding. (4) The framework is unified, effective, versatile, and efficient regarding achieving stable results under different configurations.

3.3 Multitask learning module

In prohibited item detection from X-ray images, prohibited items are likely to be blocked by other objects. However, X-ray images feature distinguishing effects, clarity and visualization abilities, so blocked areas are still somewhat visible. As shown in Fig. 6a, with the characteristics of X-ray images, current methods [7, 29] use pixel gradients, such as the Canny operator(Fig. 6b) and Sobel operator (Fig. 6c), to address occlusion problems. However, edge features contain too much noisy information, such as the edge details of items other than prohibited objects, which is not conducive to detection.

Compared with noisy edge maps, the segmented mask provides only the semantic shape of the objects of interest, which can greatly reduce the interference impact [18]. Therefore, we design an MLM learning module for combining segmentation cues to improve the detection results. An example of the segmented mask is shown in Fig. 6d, and the corresponding ground-truth segmented mask is shown in Fig. 6e.

In the segmentation branch, we use the class-aware CAM activation maps to extract the segmentation information because the detection task does not provide pixel-level labeling. The activation map (see details in Fig. 5) is not as accurate as the segmented mask map since it

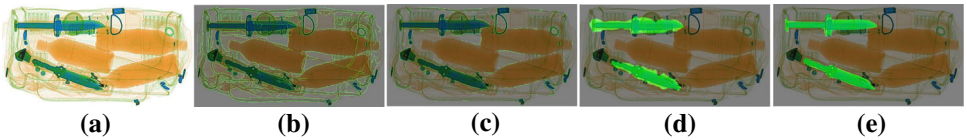


Fig. 6 Illustration of the problem with edge cues. **a** Input X-ray image. **b** Edge map obtained by the Canny operator and **c** the Sobel operator. **d** Segmented mask generated by our approach. **e** Corresponding ground truth of the segmentation obtained by manual annotation

contains more information from the foreground region than from the prohibited items. Therefore, we also employ background pseudomasks that are obtained by saliency maps [11] to alleviate the overcover in the activation map. Then, the decoder serves as the segmentor; this capability is learned by using the pseudoground-truth masks.

Any image $\mathbf{I}_j$ in the dataset is fed into the CAM to generate class-aware activation maps $\mathbf{S}_j^{dis}$; then, the semantic segmentation mask is obtained through the decoder:

$$\mathbf{O}_j = f_d(\mathbf{S}_j^{dis}, \theta), \tag{16}$$

where θ is the weight of decoder f_d, and the resolution of the output $\mathbf{O}_j$ is the same as that of input $\mathbf{I}_j$. $\mathbf{S}_j^{dis}$ is upsampled and binarized to generate the foreground pseudomasks and then combined with the background pseudomasks generated by the saliency maps to constitute the segmented mask ground truth $\mathbf{E}_j$. The loss function of the segmentation module is as follows:

$$L_{seg} = \sum_j L_{bce}(\mathbf{O}_j, \mathbf{E}_j), \tag{17}$$

where L_{bce} is the binarized cross-entropy loss function. The segmented mask $\mathbf{O}_j$ is downsampled to the same resolution as that of $\mathbf{S}_j^{dis}$, becoming $\mathbf{O}_j'$, and is fed into the classification and localization branches.

The detection branch comprises a region of interest (ROI) pooling layer, a convolutional (Conv) layer and a fully connected (FC) layer, followed by two sibling output layers. The ROI pooling layer performs dynamic max pooling over 26×26 output bins for each box. The Conv layer with a 3×3 kernel extracts abstract features. The FC layer reduces the channel number from 256 to 64. Two sibling output layers follow, that is, a scoring layer and a bounding box regression layer. Assume that N is the number of prohibited item classes, the scoring layer outputs the $(N+1)$-D vector (1 for background) representing the possibility of existence for all kinds of prohibited items, and the bounding box regression layer computes 4-D box offsets (center, width, and height). We employ the cross-entropy loss L_{ce} for classification and the L_1 loss for localization.

The final loss of the multitask learning is constructed as follows:

$$L_{det} = \sum_j (L_{ce}(\mathbf{r}_j, \mathbf{t}_j) + \beta L_1(\mathbf{p}_j, \mathbf{d}_j)), \tag{18}$$

where $\mathbf{r}_j$ and $\mathbf{t}_j$ are the prediction and ground truth for the classification, respectively, and $\mathbf{p}_j$ and $\mathbf{d}_j$ are the corresponding counterparts for localization. β is the hybrid balance factor.

The use of MLM to construct segmentation pseudomask-assisted detection has the following advantages: (1) Dual branches are used to perform different tasks, which optimizes the solution from multiple aspects; (2) Segmentation masks that accurately reflect the object category and location assist in pixel-level understanding and improve the detection accuracy.

4 Results

In this section, we verify the performance of the proposed approach compared with that of popular approaches.

4.1 Hardware and software environment

A workstation with two Intel i7-4790 3.6 GHz central processing units (CPUs) with 64 GB memory and 4 NVIDIA GTX Titan X graphics cards are used. Our approach for demonstrating the effectiveness of the proposed method is based on PyTorch.

4.2 Datasets

We evaluate our approach on the SIXray dataset, the OPIXray dataset, and the HIXray dataset.

The SIXray dataset contains 1,059,231 X-ray images collected from multiple subway stations. Prohibited items include guns, knives, wrenches, pliers, scissors, and hammers in 6 categories. The hammer class is removed in this experiment since there are fewer than 60 images containing hammers. The average size of all images is 100 K pixels, and different material objects are displayed in different colors. The dataset is divided into three subdatasets, namely, SIXray10, SIXray100 and SIXray1000, and the corresponding numbers indicate the ratio of negative samples to positive samples. Because the ratio of positive and negative samples in the SIXray100 dataset is close to

the true distribution, SIXray100 is used as the dataset for our experiments. The training set contains 7143 positive samples and 714,300 negative samples, the validation set contains 893 positive samples and 89,300 negative samples, and the test set contains 893 positive samples and 89,300 negative samples.

The OPIXray dataset contains 8885 X-ray images collected from security inspection machines at international airports, including 7109 images in the training set and 1776 images in the test set. All images vary in size. There are 5 types of prohibited items: folding knives, straight knives, scissors, utility knives, and multifunction knives. All images in this dataset contain prohibited objects, and 3 types of prohibited object occlusion levels are defined. All samples were manually marked by professional inspectors at an international airport.

The HIXray dataset contains 102,928 X-ray images collected from multiple international airports, including 82,452 images in the training set and 20,476 images in the test set. Prohibited items include portable chargers, mobile water bottles, laptops, mobile phones, tablets, cosmetics, and metallic-lighters (abbreviated as PO1, PO2, WA, LA, MP, TA, CO and ML) in 8 categories. It has high-quality images, multiple objects of interest per image, and object occlusion.

4.3 Evaluation criteria

We use evaluation criteria that others have employed and released in their work to compare our approach to popular approaches on the same datasets. We use the mAP at an intersection over union (IoU) threshold of 50% as the evaluation measure criteria for the SIXray100 and OPIXray100 datasets. All detected images are sorted according to the confidence of the detected items, and the average precision is calculated.

4.4 Implementation details

For implementation, the size of all images is adjusted to 1200×1000 resolution to meet the input requirements of the FC layer. For the SIM, the number of residual blocks is set to 5. The channel numbers of the feature maps of each scale are 64, 256, 512, 1024, and 2048. The scale interaction is iterated for 2 rounds. Then, the feature maps are fed into category-aware full convolution and global average pooling to obtain the category score vectors. In the CAM, the number of object categories in the SIXray dataset is 7 (including the background), and that in the OPIXray dataset is 6. In addition, the weight α of the cross-image attention classification loss function is set to 0.01. In the MLM, the numbers of feature map channels of the decoder are 1024, 512, 256, 64, and E (E is the number of item categories in the dataset). The numbers of channels of the two FC layers are 128 and D(D is 10 in both the SIXray dataset and the OPIXray dataset). The parameter β is 0.1 (determined by a grid search). The entire network is trained using the stochastic gradient descent algorithm, the momentum parameter is 0.9, and the weight decay coefficient is 0.007. The learning rate is 0.005 for the first 45,000 iterations and then automatically decreases according to the feedback results in the validation set. The batch size is set to 6. The numbers of epochs are set to 150 for the SIXray dataset and 120 for the OPIXray dataset.

4.5 Ablation study

We conduct extensive ablation studies to evaluate the effects of several contributions in our approach. These comparisons are performed only on the OPIXray dataset.

The backbone in the SIM We evaluate the model size, Giga floating point operations (GFLOPs), and effectiveness of different backbones, and the results are reported in Table 1. The swin transformer [12] obtains an obvious improvement owing to its natural hierarchical representation. We choose to use the swin transformer as the backbone in the following experiments.

Parameters Grid searching is employed to choose appropriate parameters in different modules. The experimental results of different numbers of scale interactions in SIM are shown in Fig. 7a. The x-axis is the interaction number, and the y-axis is the mAP. Two interactions are conducive to model decision-making, and more interactions only lead to overfitting.

The experimental results of different CFC numbers in the CAM are shown in Fig. 7b, where the x-axis is the CFC number and the y-axis is the mAP. A single CFC layer is selected to maintain a balance between effectiveness and efficiency.

The experimental results of hyperparameters in the loss function are shown in Fig. 7c. The x-axis is the parameter value, and the y-axis is the mAP. We choose $\alpha = 0.01$ and $\beta = 1.0$ according to the experimental results.

Table 1 Evaluations of different backbones on the OPIXray dataset

Module	Backbone	Model size(MB)	GFLOPs	mAP
SIM	VGG16	138.3	357.0	68.1 ± 0.3
	ResNet50	**25.5**	**97.0**	70.6 ± 0.2
	ResNet101	44.5	184.9	72.5 ± 0.2
	ResNeXt101	44.4	194.3	73.4 ± 0.3
	Swin-L	197.0	823.4	**74.5** ± 0.3

The values in bold represent the best results among different approaches

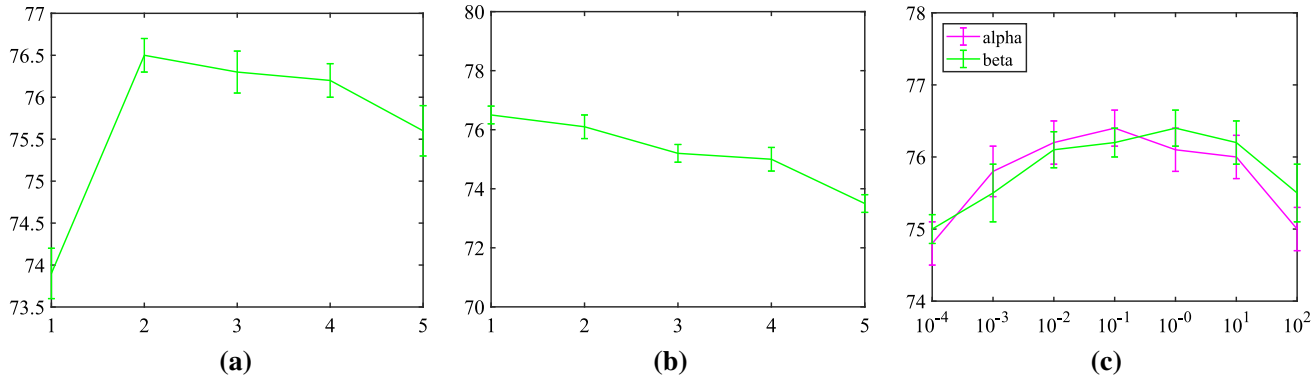


Fig. 7 Parameter selection on the OPIXray dataset. Quantitative analysis of **a** the interaction number in SIM, **b** the CFC number in CAM, and **c** hyperparameters in the loss function

The robustness of the CAM The performance varies in mAP owing to different occlusion levels, poses, and backgrounds among the corresponding reference images. Therefore, Table 2 reports the performance variation of different rounds in the experiments. The experimental results in Table 2 show that the performance is stable when 7 trials are implemented in the experiment. We also visualize common attention and discrepancy attention in the CAM module to illustrate the discrimination capability of prohibited items in X-ray images in Fig. 8. Note that input images are augmented by bounding boxes to localize prohibited items. The common attention discovers the discriminant parts of prohibited items, while normal items with similar patterns are also highlighted. Then, the discrepancy attention removes high activation regions of normal items, which makes prohibited items easy to localize.

Effectiveness Finally, we evaluate the model size, FLOPs, and effectiveness of the three proposed modules. In the benchmark model, only the residual learning encoder with the same number of layers as the proposed solution is used to extract features, and the FC layers estimate the category and location of prohibited objects. As a result, the detection performance is relatively poor. Then, the modules proposed in this paper are gradually added; the results are shown in Table 3.

Table 2 The robustness evaluations of different rounds on the OPIXray dataset

Module	Round number	mAP
CAM	3	76.6 ± 0.2
	5	77.4 ± 0.3
	7	**77.7** ± 0.2
	9	**77.7** ± 0.2

The SIM, CAM and MLM modules achieve improvements of 1.7, 1.6 and 1.4 in the mAP, respectively. This shows that our method can effectively aggregate context information, thereby improving the detection performance.

4.6 Evaluation on the SIXray dataset

This section compares the proposed method with other prohibited item detection methods. The experimental results are shown in Table 4. RetinaNet and CHR partially alleviate the occlusion problem by introducing different weights for each sample, achieving limited effects in X-ray image detection. The CHR and DetectoRS fuse features with details or semantics (from different scales) to improve the localization accuracy. The nonlocal network, swin transformer, and focal transformer explore useful context information by using a self-attention mechanism, which is helpful to detect occluded items. The DOAM and CST methods use edge information to guide the localization and are easily affected by noise factors. In contrast, our method eliminates the negative effects of high-frequency noise by generating high-level segmentation masks, gaining a 1.47% improvement in mAP.

The detection results of different methods based on the SIXray dataset are shown in Fig. 9, where the DOAM method, our method, and the ground truth are represented by blue, red and green rectangles, respectively. In Fig. 9a, our method obtains robust detection results for a variety of occlusion levels of prohibited objects. Compared with DOAM, which uses noisy edge-assisted detection, our method uses semantic segmentation information to assist in prohibited object detection.

However, our method still yields some inaccurate results, as shown in Fig. 9b. Part of the reason for this is the complex background in the X-ray images (multiple items overlap with each other), as the unique regions of the prohibited items are not correctly understood, and the information of prohibited items (such as knives) in the real

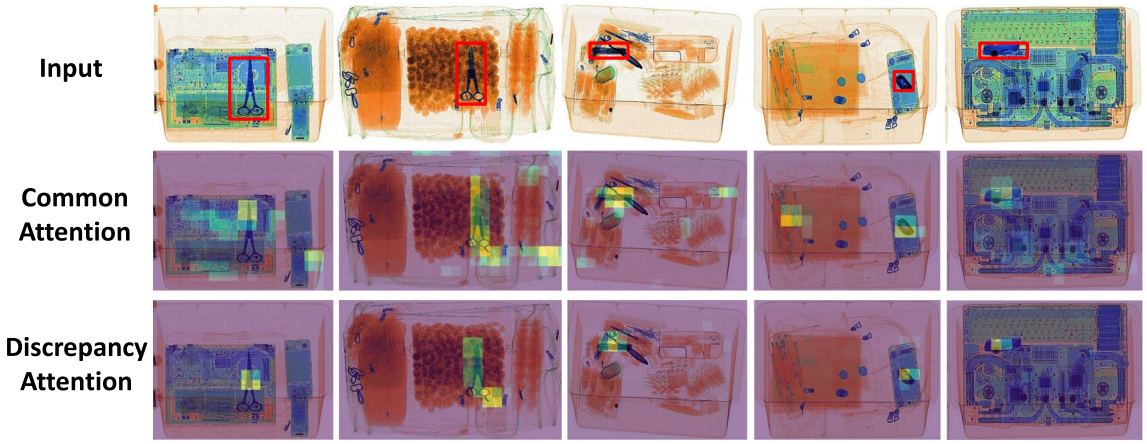


Fig. 8 Visualization of the variation of feature maps in the CAM module. Images are from the OPIXray dataset. Input images, common attention, and discrepancy attention are illustrated. Note that input images are augmented by bounding boxes to localize prohibited items

Table 3 Effectiveness evaluations of each module on the OPIXray dataset

SIM	CAM	MLM	Model size(MB)	GFLOPs	mAP
			197.0	**823.4**	74.5
✓			231.4	934.0	76.2
	✓		208.3	867.1	76.1
		✓	203.5	845.3	75.9
✓		✓	242.9	955.9	76.8
	✓	✓	212.8	889.0	76.8
✓	✓		240.7	977.7	77.3
✓	✓	✓	245.2	998.6	**77.7**

The values in bold represent the best results among different approaches

scene is limited because prohibited items occupy only a small portion of the images.

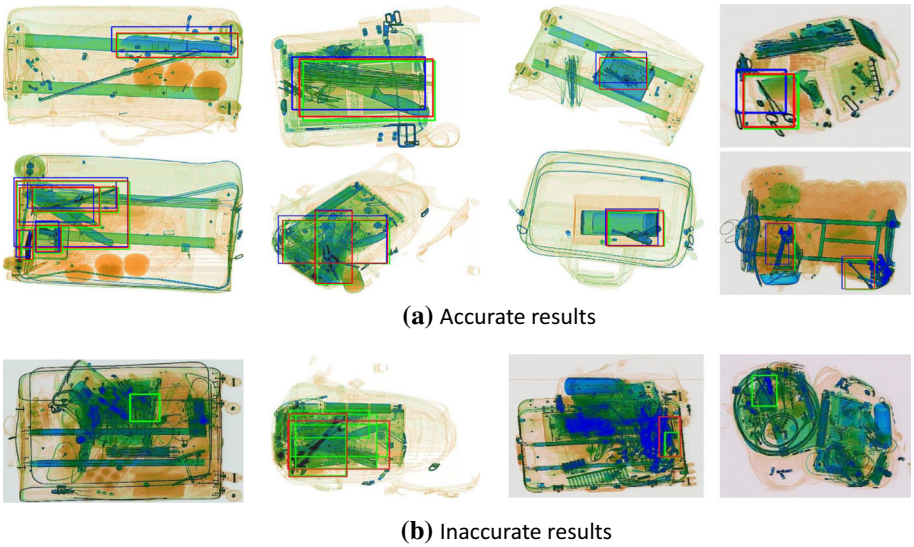
4.7 Evaluation on the OPIXray dataset

In this section, we verify the effectiveness of various methods on the OPIXray dataset. The results are shown in Table 5. FO, ST, SC, UT and MU in the table represent the folding knife, straight knife, scissors, utility knife and multifunction knife, respectively. The high-frequency noise that is generated in the edge extraction makes the CST and DOAM methods only reach 2.04% and 3.10% mAP improvements. The nonlocal network, swin transformer, and focal transformer explore nonlocal dependencies by using different structures of local regions to enhance contextual features, obtaining mAP improvements of 0.5%, 4.13%, and 5.34%, respectively. On the basis of the swin transformer and DOAM methods, our approach performs multiscale analysis with the interactions among adjacent scales, discovers semantic regions via image comparison, and uses two branches for multitask learning, which leads to an additional mAP increase of approximately 1.45%.

Table 4 Comparison of different aggregation methods on the SIXray dataset. The evaluation metric is the mAP. '*' denotes the approach we reimplemented

Method	Gun	Knife	Wrench	Pliers	Scissors	Average
RetinaNet [3]	81.16	77.27	32.44	66.87	22.61	56.07
Nonlocal [28]	82.49	78.64	33.96	67.95	23.82	57.38
CHR [13]	82.06	78.75	43.22	66.75	28.80	59.92
DetectoRS [14]	82.39	78.64	43.68	66.77	28.74	60.05
S-YOLOv4[27]	82.67	78.77	43.45	67.26	29.04	60.26
DOAM [29]*	82.55	79.05	43.63	67.14	29.21	60.33
Copy-Paste [5]	83.29	79.46	44.04	67.89	29.50	60.85
Swin Trans.[12]	83.48	79.88	44.33	67.67	30.32	61.14
CST [7]*	83.42	80.12	44.57	68.11	30.16	61.28
Focal Trans.[30]	84.27	81.33	45.36	69.40	31.19	62.31
Ours	**86.01**	**82.69**	**47.17**	**70.71**	**32.75**	**63.78**

The values in bold represent the best results among different approaches



(a) Accurate results

(b) Inaccurate results

Fig. 9 Experimental results on the SIXray dataset, where the DOAM approach, our approach, and the corresponding ground truth are illustrated in blue, red, and green bounding boxes, respectively. **a** Accurate detection results. The 1st (top) row shows the cases where the prohibited items have no or slight occlusion, and the 2nd row shows the situations where the prohibited items exhibit partial occlusions. **b** Inaccurate detection results

Table 5 Experimental comparison of the mAP on the OPIXray dataset

Method	FO	ST	SC	UT	MU	Average
RetinaNet [3]	77.07	36.06	94.61	64.64	82.15	70.91
Nonlocal [28]	77.55	36.38	95.26	64.86	82.98	71.41
CHR [13]	80.42	40.55	94.17	67.11	82.48	72.95
DetectoRS[14]	81.01	41.03	94.64	68.19	83.41	73.66
S-YOLOv4[27]	81.44	41.07	94.70	68.25	83.67	73.83
DOAM [29]*	81.37	41.50	95.12	68.21	83.83	74.01
Copy-Paste[5]	81.15	42.44	95.13	68.93	84.06	74.38
Swin Trans.[12]	82.14	42.77	95.75	69.60	84.84	75.04
CST [7]*	82.28	42.81	95.80	69.73	84.95	75.13
Focal Trans.[30]	82.96	45.13	96.35	70.92	86.02	76.25
Ours	**83.04**	**48.73**	**96.54**	**73.19**	**87.03**	**77.70**

The values in bold represent the best results among different approaches

FO, ST, SC, UT and MU represent a folding knife, a straight knife, scissors, a utility knife and a multitool knife, respectively. '*' denotes the approach we reimplemented

The accurate detection results in Fig. 10a show that our method is robust to variance in the input. Fig. 10b shows some of the inaccurate detection results. Note that because the useful information and the interference information of the complex background are intertwined, the serious occlusions caused by other objects greatly affect the performance. In addition, other factors, such as the camera view, inevitably expand the intravariance. For example, a straight knife becomes thinner at a specific observation spot, drifting from the typical characteristics of the prohibited item. Multiview analysis may be a potential solution, collecting observed cues from different views for learning.

4.8 Evaluation on the HIXray dataset

We also conduct experiments on the HIXray dataset. The experimental results are reported in Table 6, and some detection results are illustrated in Fig. 11. PO1, PO2, WA, LA, MP, TA, CO, and ML represent portable chargers 1 (lithium-ion prismatic cell), portable chargers 2 (lithium-ion cylindrical cell), mobile water bottles, laptops, mobile phones, tablets, cosmetics, and metallic lighters, respectively. Note that our results based on the HIXray dataset yield conclusions similar to those from our results on the OPIXray dataset. Our method obtains 83.21% in mAP, values which are 1.11% more than the runner-up.

In Fig. 11, the detected outputs of the DOAM and our approach and the corresponding ground truth are illustrated. The accurately detected bounding boxes in Fig. 11a certify that our approach is robust to variations in background clutter. Feature maps obtained from multiple perception fields and analyzed by common and discrepancy knowledge across different images help to locate positions and recognize categories of prohibited items.

Fig. 11b also illustrates the inaccurate inference results. Note that items in 'cosmetic' are sometimes missed owing to its diversity in both appearance and shape. Curriculum learning [6] is a potential solution, because it learns the pattern of objects from general to specific in a cascade manner, which partially solves the cosmetic diversity problem.

4.9 Discussion

If the prohibited item is partially occluded, the information from the unoccluded distal part can be employed for discrimination by using multiple scale perceptions. In multi-scale analysis, CHR [13] and DetectoRS [14] deliver only

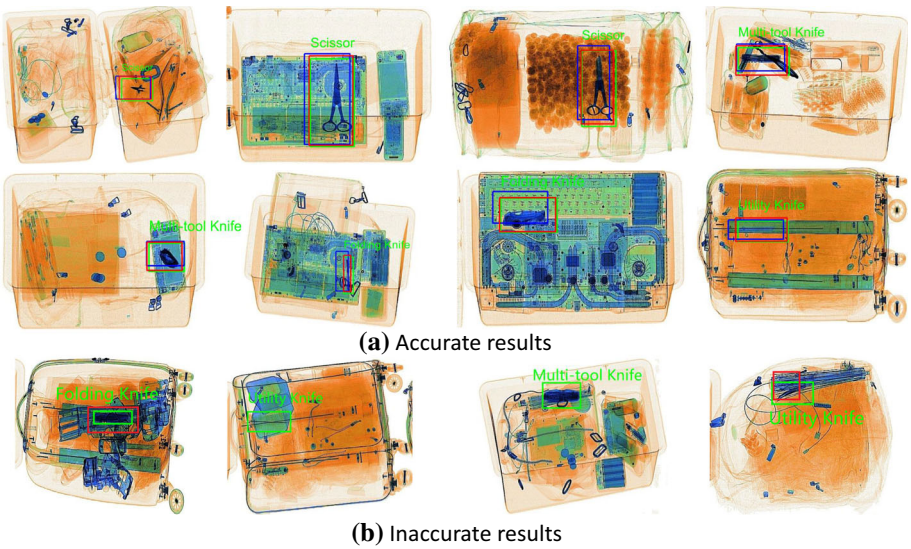


Fig. 10 Experimental results on the OPIXray dataset, where the DOAM approach, our approach, and the corresponding ground truth are illustrated in blue, red, and green bounding boxes, respectively. **a** Accurate detection results. **b** Inaccurate detection results

Table 6 Experimental comparison of the mAP on the HIXray dataset

Method	PO1	PO2	WA	LA	MP	TA	CO	ML	Average
RetinaNet [3]	88.73	86.31	86.71	89.82	88.82	88.81	63.44	13.35	75.74
Nonlocal [28]	88.82	87.73	87.61	89.73	89.51	88.66	63.75	12.92	76.23
CHR [13]	88.67	88.52	88.52	90.23	89.44	89.43	69.61	14.42	77.32
DetectoRS[14]	90.84	88.43	88.83	91.94	90.93	90.92	65.53	15.44	77.85
S-YOLOv4[27]	91.04	89.91	89.83	91.91	91.74	90.82	65.44	15.63	78.44
DOAM [29]*	93.22	92.60	90.56	95.65	96.04	92.41	61.79	14.38	79.61
Copy-Paste[5]	93.84	92.75	91.63	94.71	93.57	92.63	64.71	16.90	80.28
Swin Trans.[12]	94.82	93.88	92.44	97.49	97.15	94.72	63.94	15.14	81.22
CST [7]*	95.33	94.71	92.72	97.81	98.22	94.53	63.91	16.52	81.72
Focal Trans.[30]	95.73	94.72	93.32	98.31	98.01	95.67	64.88	16.02	82.10
Ours	**96.11**	**95.02**	**93.94**	**98.32**	**98.53**	**95.80**	**65.62**	**19.97**	**83.21**

The values in bold represent the best results among different approaches

PO1, PO2, WA, LA, MP, TA, CO, and ML represent portable chargers 1 (lithium-ion prismatic cell), portable chargers 2 (lithium-ion cylindrical cell), mobile water bottles, laptops, mobile phones, tablets, cosmetics, and metallic lighters, respectively. '*' means the approach we reimplemented

high-level visual cues to assist midlevel features and achieve limited improvement in object localization. To handle this predicament, our SIM module also incorporates detailed information from low-level feature maps and bidirectional mining of contextual semantic information to improve localization accuracy. The feature interaction of neighboring scales enhances the feature representation, alleviates the semantic gaps between features in different perception domains, and allows the model to obtain a comprehensive understanding of the entirety and different parts of the same object. It can also effectively capture the appearance variations caused by severe occlusion, where the visual features are seriously insufficient.

The rich contextual semantic information between images can be explored to understand the unique parts of prohibited items. The common attention in the CAM module explores cross-image shared semantics, which helps the classifier to proficiently perceive the common semantic labels over the coattentive regions. Discrepancy attention in the CAM module focuses on unshared semantics, which enables the classifier to capably separate the semantic patterns of different objects. Actually, the cross-attention [28] or transformer [12, 30] mechanism can not only explore similarity in local and global regions but also discover and discriminate semantics by using images containing items in the same of different classes, at the cost of computational complexity.

Although the segmentation-based branch (our MLM module) does not receive accurate pixel-level outputs, it is more robust than edge-based methods [7, 29] because an edge is affected by factors from other sides, while an object region contains much information to describe the specific

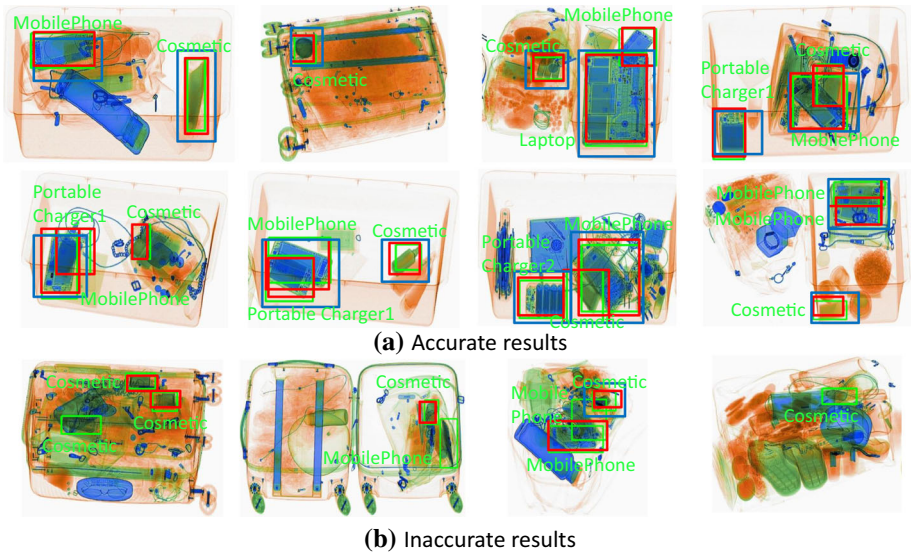


Fig. 11 Experimental results on the HIXray dataset, where the DOAM approach, our approach, and the corresponding ground truth are illustrated in blue, red, and green bounding boxes, respectively. **a** Accurate detection results. **b** Inaccurate detection results

class to which the object belongs, which reflects the object category and location assisting in pixel-level understanding and improving the detection accuracy. In addition, dual branches are used to perform different tasks (detection and segmentation), which optimizes the shared feature maps from both the pixel level and object level.

The basis of our approach is that the occluded part is still partially visible in the X-ray image. Therefore, we learn the multiscale analysis, characteristics, and segmentation of the interacting part to improve the perception of prohibited items. In general object detection, the occluded part is totally invisible; as a result, our approach does not work in this situation owing to the absence of valuable partially observed appearance information.

In future work, we will explore the relationship between different viewpoints or depths through multiview- or computed tomography-based approaches to expand the method and address the challenges under a single viewpoint, which will further improve the effectiveness of prohibited item detection from X-ray images. Curriculum learning [6], learning the pattern of objects from general to specific in a cascade manner, is also a potential solution to handle the problem of diversity in prohibited item detection.

5 Conclusion

Here, we present a new method that employs multilayer feature interaction to improve the perception ability of the model. The proposed cross-image analysis can learn the pixel-level semantics of objects in a weakly supervised manner. This pixel-level information can further assist in prohibited item detection, especially in the case of missing information, such as from occlusion. Experimental results on the SIXray dataset, the OPIXray dataset, and the HIXray dataset show that our approach outperforms other popular approaches by margins of 1.47%, 1.45%, and 1.11% in mAP.

Acknowledgements The authors would like to thank AJE (www.aje.com) for its linguistic assistance during the preparation of this manuscript.

Declarations

Conflict of interest All authors declare that they have no conflicts of interest regarding the publication of this paper.

References

1. Akcay S, Atapour-Abarghouei A, Breckon TP (2018) Ganomaly: semi-supervised anomaly detection via adversarial training. In: Asian conference on computer vision, pp 622–637
2. Akcay S, Kundegorski ME, Willcocks CG et al (2018) Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. IEEE Trans Inf Forens Security 13(9):2203–2215
3. Cui Y, Oztan B (2019) Automated firearms detection in cargo x-ray images using retinanet. In: Anomaly detection and imaging with X-Rays (ADIX) IV, p 109990P
4. Fu J, Liu J, Tian H et al (2019) Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3146–3154
5. Ghiasi G, Cui Y, Srinivas A et al (2021) Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2918–2928
6. Hacohen G, Weinshall D (2019) On the power of curriculum learning in training deep networks. In: International conference on machine learning, pp 2535–2544
7. Hassan T, Akcay S, Bennamoun M et al (2020) Cascaded structure tensor framework for robust identification of heavily

occluded baggage items from x-ray scans. arXiv preprint arXiv:2004.06780
8. He K, Zhang X, Ren S, et al (2016) Identity mappings in deep residual networks. In: Proceedings of the european conference on computer vision, pp 630–645
9. Jain DK et al (2019) An evaluation of deep learning based object detection strategies for threat object detection in baggage security imagery. Pattern Recogn Lett 120:112–119
10. Lin TY, Goyal P, Girshick R, et al (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
11. Liu JJ, Hou Q, Cheng MM et al (2019) A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3917–3926
12. Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
13. Miao C, Xie L, Wan F, et al (2019) Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2119–2128
14. Qiao S, Chen LC, Yuille A (2021) Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10213–10224
15. Ren S, He K, Girshick R et al (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Proc Adv Neural Inf Process Syst 28:91–99
16. Tao R, Wei Y, Jiang X, Li H, Qin H, Wang J, Ma Y, Zhang L, Liu X (2021) Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10923–10932
17. Tian Y, Chen T, Cheng G et al (2022) Global context assisted structure-aware vehicle retrieval. IEEE Trans Intell Transp Syst 22(12):1–10
18. Tian Y, Cheng G, Gelernter J et al (2020) Joint temporal context exploitation and active learning for video segmentation. Pattern Recogn 100:107158
19. Tian Y, Gelernter J, Wang X et al (2018) Lane marking detection via deep convolutional neural network. Neurocomputing 280:46–55
20. Tian Y, Gelernter J, Wang X et al (2019) Traffic sign detection using a multi-scale recurrent attention network. IEEE Trans Intell Transp Syst 20(12):4466–4475
21. Tian Y, Hu W, Jiang H et al (2019) Densely connected attentional pyramid residual network for human pose estimation. Neurocomputing 347:13–23
22. Tian Y, Wang H, Wang X (2017) Object localization via evaluation multi-task learning. Neurocomputing 253:34–41
23. Tian Y, Wang X, Wu J et al (2019) Multi-scale hierarchical residual network for dense captioning. J Artif Intell Res 64:181–196
24. Tian Y, Zhang Y, Xu H et al (2022) 3d tooth instance segmentation learning objectness and affinity in point cloud. ACM Trans Multimed Comput Commun Appl 18:202–211
25. Tian Y, Zhang Y, Zhou D et al (2020) Triple attention network for video segmentation. Neurocomputing 417:202–211
26. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Proceedings of the advances in neural information processing systems, pp 5998–6008
27. Wang CY, Bochkovskiy A, Liao HYM (2021) Scaled-yolov4: Scaling cross stage partial network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13029–13038
28. Wang X, Girshick R, Gupta A, et al. (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
29. Wei Y, Tao R, Wu Z, et al. (2020) Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In: Proceedings of the ACM international conference on multimedia, pp 138–146
30. Yang J, Li C, Zhang P, et al. (2020) Focal self-attention for local-global interactions in vision transformers. In: Proceedings of the advances in neural information processing systems, pp 138–146
31. Yu J, Lin Z, Yang J, et al. (2019) Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4471–4480