PAPER Special Section on Mathematical Systems Science and its Applications

Prohibited Item Detection Within X-Ray Security Inspection Images Based on an Improved Cascade Network

Qingqi ZHANG[†], **Xiaoan BAO**[††], ***Nonmembers***, **Ren WU**[†††], **Mitsuru NAKATA**[††††], ***and*** **Qi-Wei GE**[†††††a)], ***Members***

SUMMARY Automatic detection of prohibited items is vital in helping security staff be more efficient while improving the public safety index. However, prohibited item detection within X-ray security inspection images is limited by various factors, including the imbalance distribution of categories, diversity of prohibited item scales, and overlap between items. In this paper, we propose to leverage the Poisson blending algorithm with the Canny edge operator to alleviate the imbalance distribution of categories maximally in the X-ray images dataset. Based on this, we improve the cascade network to deal with the other two difficulties. To address the prohibited scale diversity problem, we propose the Re-BiFPN feature fusion method, which includes a coordinate attention atrous spatial pyramid pooling (CA-ASPP) module and a recursive connection. The CA-ASPP module can implicitly extract direction-aware and position-aware information from the feature map. The recursive connection feeds the CA-ASPP module processed multi-scale feature map to the bottom-up backbone layer for further multi-scale feature extraction. In addition, a Rep-CIoU loss function is designed to address the overlapping problem in X-ray images. Extensive experimental results demonstrate that our method can successfully identify ten types of prohibited items, such as Knives, Scissors, Pressure, etc. and achieves 83.4% of mAP, which is 3.8% superior to the original cascade network. Moreover, our method outperforms other mainstream methods by a significant margin.

key words: *X-ray security inspection images, prohibited item detection, image fusion, multi-scale feature extraction*

1. Introduction

As urban populations and crowd densities at public transportation hubs grow, security inspection is becoming increasingly important in protecting public safety [1]. Security inspection machine is the most widely used security inspection equipment [2]. It uses X-ray technology to scan a traveler's package and generate an irradiation image in real time. Currently, most of the work of security inspection still relies on highly trained security staff to carefully identify by eye whether there are any prohibited items in the irradiation image [3], [4]. As security staff fulfills a demanding occupation, being in a high-pressure work environment for elongated periods may cause false detection or missed detection of prohibited items, which may seriously threaten public safety [5]. Moreover, frequent shift changes consume many human resources and increase labor costs.

With the substantial development of artificial intelligence technologies, automatic security inspection of prohibited items has become possible in recent years. Machine learning and deep learning algorithms are the main methods for prohibited item detection within X-ray security inspection images. Muhammet et al. [6] utilize the Bag of Visual Word (BoVW) framework with SVMs structure to classify prohibited items. Mery et al. [7] proposed to use a method based on multiple X-ray views to detect regular prohibited items with very defined shapes and sizes. The main drawback of these machine learning approaches is the reliance on hand-crafted features that require manual engineering. Wang et al. [8] and Miao et al. [9] proposed a selective dense attention network and class-balanced hierarchical refinement (CHR) approach, respectively. These deep learning methods have achieved better performance compared to machine learning methods.

However, three challenges still appeal to us in the prohibited items detection task. First, the X-ray security inspection dataset has an imbalanced distribution of categories. Deep learning as a standard data-driven technique, a balanced distribution of categories in the dataset is the cornerstone of the algorithm to achieve better performance. The dataset used in this paper, as shown in Fig. 1, consists of two parts provided by iFLYTEK CO.LTD: X-ray images of the entire package and X-ray images of a separate prohibited item. To alleviate the category imbalance problem, we propose to leverage the Poisson blending algorithm with the Canny edge operator to fuse an X-ray image of a separate prohibited item with an X-ray image of the entire package. This data enhancement method can naturally fuse the two X-ray images with minimal noise and increase the diversity and complexity of the samples.

Second, diversity of prohibited item scales. The size of prohibited items in the same X-ray image varies. There is also variation in the size of the same type of prohibited items in different X-ray images. To address the challenge of detecting prohibited items at diverse scales, we propose an approach named Re-BiFPN. For comparison, a cascaded network [10] merges multiple detectors and leverages FPN [11]

Manuscript received April 14, 2023.
Manuscript revised October 10, 2023.
Manuscript publicized January 16, 2024.

[†]The author is with the Graduate School of East Asian Studies, Yamaguchi University, Yamaguchi-shi, 753-8540 Japan.
[††]The author is with the School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, 310018 China.
[†††]The author is with the Faculty of Information Science, Shunan University, Shunan-shi, 745-8566 Japan.
[††††]The author is with the Faculty of Education, Yamaguchi University, Yamaguchi-shi, 753-8513 Japan.
[†††††]The author is with the Yamaguchi University, Yamaguchi-shi, 753-8511 Japan.
a) E-mail: gqw@yamaguchi-u.ac.jp
DOI: 10.1587/transfun.2023MAP0007

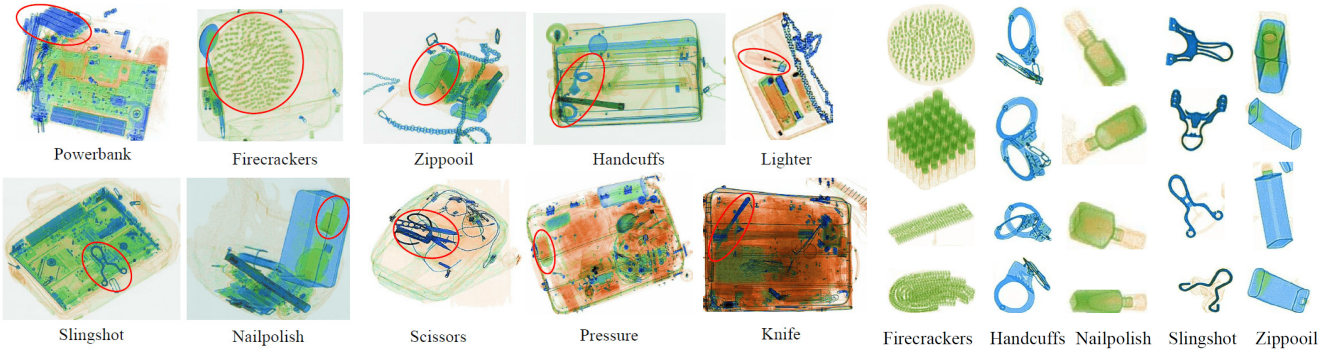


Fig. 1 The visualization examples of the X-ray security inspection dataset.

for feature extraction. Unlike the FPN, which mainly concentrates on managing multi-scale features through straightforward aggregation, Re-BiFPN presents a novel theoretical advancement. It establishes a recursive multi-scale structure and incorporates the coordinate information of prohibited items into the feature layers. This unique design allows our model to refine multi-scale information iteratively, enhancing multi-scale representations. Moreover, it equips the model with the ability to discern the relative positions and spatial relationships among prohibited items across different scales.

Third, the problem of overlapping prohibited items has been receiving the attention of most researchers, such as [12]–[14]. In prohibited item detection, however, no loss function is designed for this problem. Since the evaluation metric for the prohibited item detection task is IoU (Intersection over Union), the loss function in the original cascade network, which calculates the loss of the prediction box's four points, is unsuitable for this task. We designed a new loss function, Rep-CIoU, to make the model more robust to overlapping items, which considers the IoU between multiple prediction boxes and the centroid distance between the prediction box and ground-truth. It can effectively prevent multiple prediction boxes from filtering out by NMS (Non-Maximum Suppression) when the IoU generated by a particular prediction box and other surrounding prediction boxes is large or their centroids' distance is small.

In summary, the contributions of this paper are as follows: (1) We propose to utilize the Poisson blending algorithm with the Canny edge operator to fuse an X-ray image of a separate prohibited item with an X-ray image of the entire package, which can naturally fuse the two X-ray images with minimal noise and increase the diversity and complexity of the samples. (2) We propose the Re-BiFPN feature fusion method, which includes a CA-ASPP module and a recursive connection. The method can learn the coordinate information implicit in the feature maps while improving the network's ability to extract multi-scale features. (3) We designed a new loss function, Rep-CIoU, to make the model more robust to overlapping items, which considers the IoU between multiple prediction boxes and the centroid distance between the prediction box and ground-truth. This loss function can effectively reduce missed detection due to overlap.

2. Related Works

In this section, we introduce previous related works that use machine learning as well as deep learning algorithms to detect prohibited items.

2.1 Machine Learning-Based Detection Methods

As machine learning-based methods, Muhammet et al. [6] suggested using the BoVW (Bag of Visual Word) framework combined with the SVM algorithm to detect prohibited items. Mery et al. [7] proposed a method based on multiple X-ray views to detect regular prohibited items. The method consists of two steps: "structure estimation", to obtain a geometric model of the multiple views from the object to be inspected (baggage); and "parts detection", to detect the parts of interest (prohibited items). Inspired by [6] and the advantages of neural networks, Akcay et al. [15] employed a transfer learning paradigm combined with an SVM such that a pre-trained CNN can be optimized explicitly as a later secondary process that targets this specific application domain. Roomi et al. [16] trained fuzzy KNN classifiers were trained with contextual descriptors and Zernike polynomials to study pistol detection, but only fifteen image examples were evaluated.

However, these methods are designed mainly for image classification and hence have weak ability to extract feature of X-ray images.

2.2 Deep Learning-Based Detection Methods

With the development of artificial intelligence, it has become possible to apply deep learning algorithms to prohibited item detection. In addition to machine learning algorithms [15], Akcay et al. [17] also studied deep learning strategies to improve the performance of cluttered datasets further. They explored the applicability of multiple CNN-driven detection paradigms and illustrated the comparative performance of these techniques, including sliding window-based CNN, Faster region-based CNNs and region-based fully convolutional networks. Wang et al. [8] collected a dataset named PIDray and proposed a selective dense attention network

consisting of a dense attention module and a dependency refinement module. The dense attention module is used to capture the discriminative features, and the dependency refinement module is constructed to exploit the dependencies among multi-scale features. Miao et al. [9] collected a dataset named SIX-ray and presented a CHR model, which achieves class balance through a class-balanced loss function. The CHR model achieves a remarkable detection advantage on the dataset with few positive training samples. Nevertheless, these approaches need pay more attention to the X-ray images dataset's category imbalance problem.

To learn the different scales of prohibited items, Zhang et al. [18] proposed a novel asymmetrical convolution multi-view neural network (ACMNet) that includes an asymmetrical tiny convolution module, a detailed convolution multi-view module, and a fusion strategy of the multi-scale feature map. However, from their experimental results, the detection accuracy of some targets was not significantly improved. The fundamental reason is that there is a significant semantic gap between each feature layer. Moreover, there is a lack of handling prohibited item coordinate information across different scale feature maps. Feature pyramids are mainly used to improve the semantic gap in target detection [19]–[22]. However, unlike typical feature pyramid methods [23], [24] designed for color images, our Re-BiFPN is tailored for X-ray security datasets, which predominantly contain monochrome X-ray images, and it constructs a recursive multi-scale structure and possesses sensitivity to the coordinates of prohibited items within these images.

The loss function for the target detection task consists of two parts, Classification Loss and bounding box regression Loss. The bounding box regression Loss for the target detection task has undergone the evolution of Smooth L1 Loss [25], IoU Loss [26], Repulsion Loss [27], GIoU Loss [28], DIoU Loss [29] and CIoU Loss [30] in recent years. CIoU Loss can evaluate the overlap area (i.e., IoU), centroid distance, and aspect ratio between the prediction box and the ground truth. However, for the overlap problem, CIoU Loss ignores the relationship between a particular prediction box and other prediction boxes that are close to it. As far as we know, no loss function is designed for the problem of overlapping in prohibited item detection.

Through the above analysis of these related deep learning works, we approach this prohibited item detection problem by (a) paying more attention to the X-ray images dataset's category imbalance problem, (b) achieving more efficient cross-scale connectivity and weighted feature fusion to design a feature fusion method, Re-BiFPN, and (c) designing a new loss function, Rep-CIoU, to make the model more robust to overlapping items.

3. Proposed Method

Based on the analytical results of last section, here we propose a method to detect prohibited item within X-ray security inspection images.

3.1 The Overall Framework of Our Method

Our method is a multi-stage target detection architecture consisting of a series of IoU threshold trained detectors that are continuously improved. The cascade process can continuously change the distribution of candidate boxes and resample them by adjusting the thresholds [10]. The overall framework of our method is shown in Fig. 2.

We construct the training set by fusing the X-ray images of a single prohibited item with the X-ray images of the entire package. For clarity, we have artificially marked the location of a single prohibited item in the fused image with a red circle, as shown in the "Training set" of Fig. 2. Our method employs ResNeXt-101(32x4d) [31] as the backbone network. The characters on the arrow and the white circles indicate the feature map of the backbone. The colored circles indicate the multi-scale features in our proposed Re-BiFPN. The up arrow in the Re-BiFPN indicates down-sampling; the down arrow indicates up-sampling; horizontal arrow and curved arrow indicate connection operations; the red arrow is recursive connection. The CA-ASPP is coordinate attention atrous spatial pyramid pooling module. RPN is the region proposal network. "pooling" is region-wise feature extraction. "B0" is proposals in all architectures. The "Rep-CIoU" is our proposed bounding box regression loss function. And

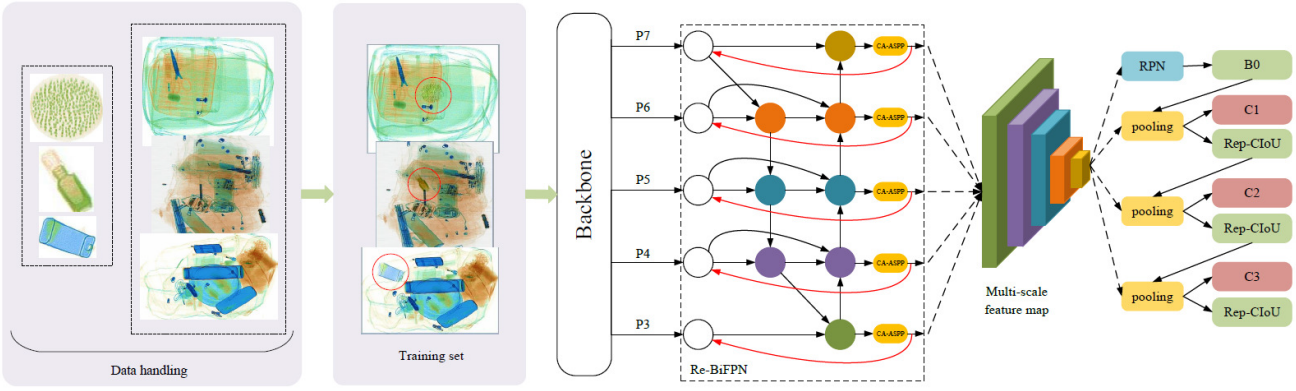


Fig. 2 The overall framework of our method.

"C" is classification loss function.

3.2 Data Handling

The issue of category imbalance refers to a situation in a training dataset where there is a significant disparity in the number of samples among different categories. This imbalance may cause the model to overly optimize for the majority category while neglecting the minority category, thereby reducing the model's generalization capability. In X-ray security inspection dataset, category imbalance may stem from the much higher occurrence rates of certain prohibited items, such as knives and lighters, compared to others. Additionally, obtaining X-ray images of specific prohibited items, like fireworks and slingshots, becomes more challenging due to their dangerous nature and rarity. To address the imbalance distribution of categories, we propose utilizing the Poisson blending algorithm [32] in conjunction with the Canny edge operator. This method aims to mitigate the category imbalance issue by fusing X-ray images of individual prohibited items with images of entire packages.

The image in Fig. 1(b) is source image for the Poisson blending. First, we leverage the excellent contour detection capability of the Canny edge operator to extract the contour information of the prohibited item, thus avoiding the introduction of out-of-contour noise. The operation on the source image is shown in Fig. 3. The source image is randomly rotated or scaling. The random rotation range is $[0, 360°]$. The range of random scaling is $\frac{1}{n}$ times the length and width of the original image, $n \in \{n|1 \leq n \leq 10, n \in Z\}$. Then contour detection is performed, and interfering parts other than the target contour is removed to obtain an image to be fused.

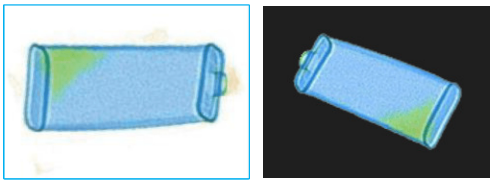
After contour detection, we utilize the Poisson blending algorithm to maximize the retention of gradient information of the image to be fused to make the fusion boundary smoother. The Poisson blending of Fig. 3(b) and Fig. 4(a) is performed. Finally, an image is obtained after Poisson blending as shown in Fig. 4(b).

Compared to the addition operation, as shown in Fig. 4(b) and Fig. 4(c), our method offers superior image fusion quality and improved edge blending effects. This advantage stems from Poisson blending's consideration of gradient discrepancies between the target and source images, enabling a more natural fusion that avoids abrupt changes in edges and colors. Conversely, the addition operation merely involves straightforward pixel value summation, which might lead to less smooth and natural image fusion outcomes.

In Fig. 4(d), we display an actual image with occlusion. Both the object marked in red in Fig. 4(d) and the blended object in Fig. 4(b) represent Zippooil. It's evident that the image generated by our method reproduces the occlusion nearly as accurately as the actual image. In addition, the image produced by our method maintains edge details similar to the actual image and prevents abrupt transitions in edges and colors.

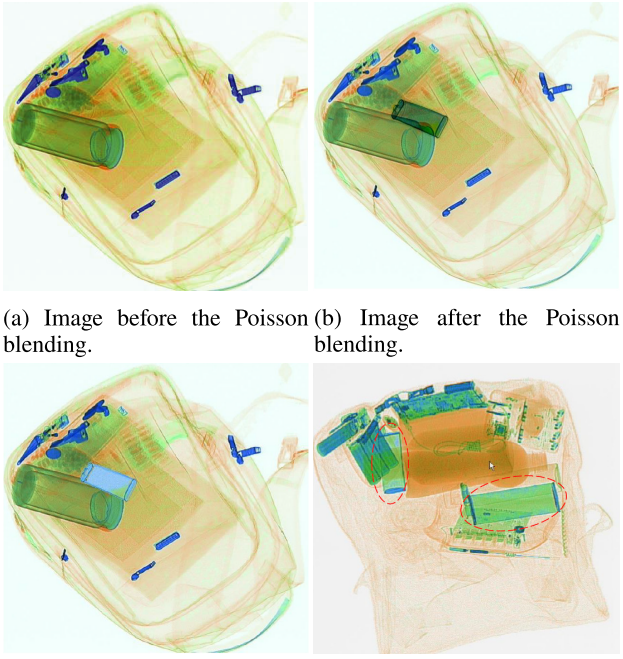
We employ the Structural Similarity Index (SSIM) to quantitatively assess the similarity and structural preservation between images before and after each Poisson blending operation, as depicted in Fig. 5. SSIM is an effective metric for evaluating image quality and is frequently used to measure the resemblance between two images in terms of pixel-level differences, structural coherence, and textures. The findings, displayed in Fig. 5, show high SSIM values. The highest SSIM value achieved is 98.82%, the lowest stands at 93.03%, and the average is 95.50%. These figures underscore a considerable similarity between the images before and after Poisson blending, especially concerning textures, structure, and intricate details.

We analyze samples both before and after the application of our method, as depicted in Fig. 6. From this figure, it becomes evident that using our method results in a substantial increase in the sample counts for five categories: Fire-



(a) Source image. (b) Image to be fused.

Fig. 3 Contour detection is performed on the source image utilizing the Canny edge operator. For the clarity of the presentation, we have marked the borders of the source image in blue. After processing by the Canny edge operator, the boundary of the image to be fused is the contour of the prohibited item.



(a) Image before the Poisson blending. (b) Image after the Poisson blending.

(c) Image obtained through addition. (d) Actual image with occlusion in the dataset.

Fig. 4 An example of image fusion utilizing the Poisson blending with the Canny edge operator. For clarity, we also show the image obtained by utilizing the general addition operation, denoted as (c), the actual image with occlusion in the dataset, denoted as (d), and compare them to (b).

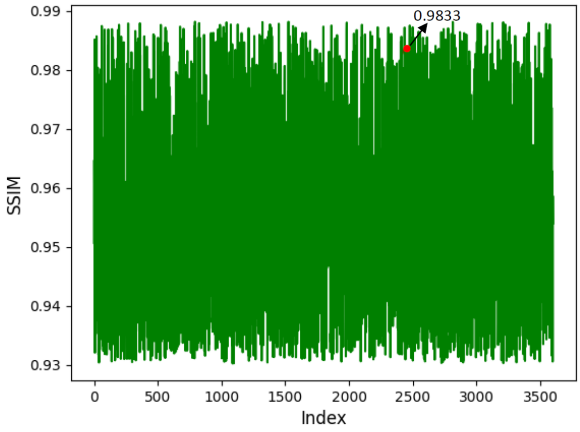
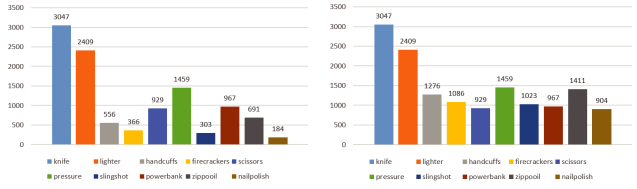


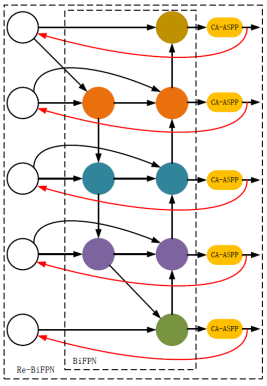
Fig. 5 Line chart of Structural Similarity Index (SSIM) for each image before and after Poisson blending. The maximum SSIM value reaches 98.82%, the minimum is 93.03%, with an average of 95.50%. The red-marked point represents the SSIM value of the images before and after Poisson blending shown in Fig. 4.



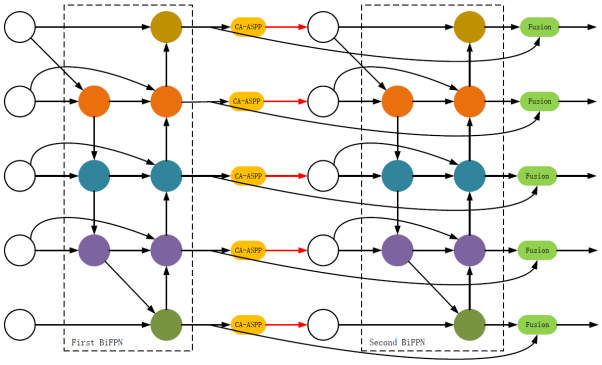
(a) Distribution of prohibited item categories before Poisson blending.

(b) Distribution of prohibited item categories after Poisson blending.

Fig. 6 Distribution of prohibited item categories before and after Poisson blending.



(a) The structure of Re-BiFPN.



(b) Unrolling Re-BiFPN.

Fig. 7 The structure of Re-BiFPN and the expanded view. The white circles represent the feature maps extracted by backbone. The colored circles indicate the multi-scale features in the Re-BiFPN structure.

crackers, Handcuffs, Nailpolish, Slingshot, and Zippooil. Moreover, aside from Knife and Lighter which are more commonly encountered and thus easier to collect in larger quantities, the sample distribution across categories appears relatively balanced. It's crucial to note that our method is only utilized during the model's training phase, while the evaluation is conducted using the original images.

3.3 Re-BiFPN Feature Fusion

To improve the network's ability to learn multi-scale features of prohibited items, we propose the Re-BiFPN feature fusion method, which achieves more efficient cross-scale connectivity and weighted feature fusion. The Re-BiFPN includes a recursive connection and a CA-ASPP module.

Figure 7(a) shows the structure of Re-BiFPN proposed in this paper. Re-BiFPN connects the layers in BiFPN to the bottom-up backbone network through additional recursive connections to form a recursive structure. The red arrow in Fig. 7(a) is the recursive connection. Specifically, this recursive connection brings the features with rich multi-scale information back to the lower-level backbone network, which is not rich enough in multi-scale information, thus enhancing the representation of features to achieve efficient cross-scale connectivity and weighted feature fusion. Figure 7(b) is the expanded view of Fig. 7(a).

The structure of CA-ASPP module is shown in Fig. 8. The CA-ASPP module takes the output features of the first BiFPN structure as input and converts them into the features used in the second bottom-up backbone network in Fig. 7(b). Simultaneously, it captures cross-channel, direction-aware, and position-sensitive information to help the model locate and identify prohibited items.

As shown in Fig. 8, in addition to $1 \times 1Conv$ and $1 \times 1Pooling$, we set up $3 \times 3Conv$ dilated convolution with the expansion ratio of 4, 8, and 12 for capturing the multi-scale information in the feature maps. And then, the two vectors are obtained by average pooling for horizontal and vertical directions, respectively. These two vectors with embedded direction-aware and position-sensitive information are encoded as two attention maps, each capturing the long-range dependencies of the input feature map along a spatial direction. The Concat operation and BN operation are performed on these two vectors. Next, the split operation is performed and the weights are obtained after the Sigmoid activation function. Finally, the weights are added to the $C \times H \times W$ feature maps.

Mathematical expression of the Re-BiFPN structure: Let P_i^{td} and P_i^{out} denote the intermediate feature layer and the output feature layer of the first BiFPN structure, re-

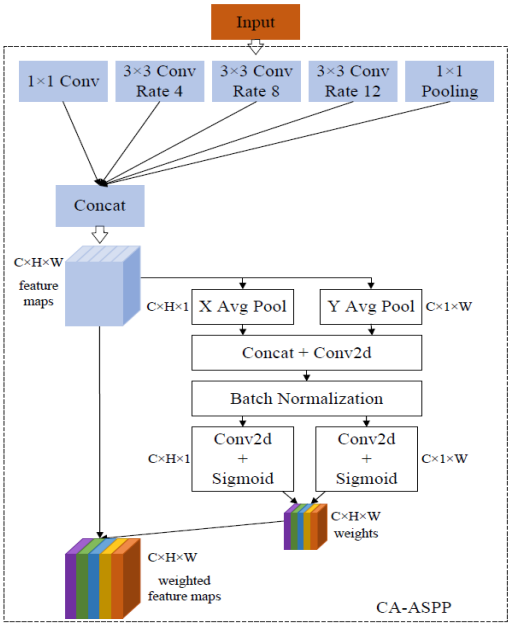


Fig. 8 The structure of CA-ASPP module.

spectively. *Resize* denotes up-sampling and down-sampling. Both w_i and $w_i^{'}$ denote learnable weights. P_i^{td} and P_i^{out} are calculated according to Eq. (1) and Eq. (2), respectively. Let R_i denote the feature transformation before connecting the features to the bottom-up backbone network. Let F_i^{td} represent the intermediate feature layer of the second BiFPN structure. Let F_i^{out} represent the output layer features of the second BiFPN structure. Then, the intermediate feature layer and output layer features of the second BiFPN structure can be derived according to Eq. (3) and Eq. (4), respectively. To prevent the divisor from being zero set ε in the formula to a small constant. The fusion module in Fig. 7(b) is used to fuse P_i^{out} and F_i^{out} together. To further improve the efficiency, the feature fusion process of Re-BiFPN uses deeply separable convolution [33].

$$P_i^{td} = \begin{cases} \frac{w_1 P_i^{in} + w_2 Resize(P_{i+1}^{td})}{w_1 + w_2 + \varepsilon} & \text{if } i = 4,5 \\ \frac{w_1 P_i^{in} + w_2 Resize(P_{i+1}^{in})}{w_1 + w_2 + \varepsilon} & \text{if } i = 6 \end{cases} \tag{1}$$

$$P_i^{out} = \begin{cases} \frac{w_1^{'} P_i^{in} + w_2^{'} Resize(P_{i+1}^{td})}{w_1^{'} + w_2^{'} + \varepsilon} & \text{if } i = 3 \\ \frac{w_1^{'} P_i^{in} + w_2^{'} P_i^{td} + w_3^{'} Resize(P_{i-1}^{out})}{w_1^{'} + w_2^{'} + w_3^{'} + \varepsilon} & \text{if } i = 4,5,6 \\ \frac{w_1^{'} P_i^{in} + w_3^{'} Resize(P_{i-1}^{out})}{w_1^{'} + w_3^{'} + \varepsilon} & \text{if } i = 7 \end{cases} \tag{2}$$

$$F_i^{td} = \begin{cases} \frac{w_1 R_i(P_i^{out}) + w_2 Resize(R_i(F_{i+1}^{td}))}{w_1 + w_2 + \varepsilon} & \text{if } i = 4,5 \\ \frac{w_1 R_i(P_i^{out}) + w_2 Resize(R_i(P_{i+1}^{out}))}{w_1 + w_2 + \varepsilon} & \text{if } i = 6 \end{cases} \tag{3}$$

$$F_i^{out} = \begin{cases} \frac{w_1^{'} R_i(P_i^{out}) + w_2^{'} Resize(F_{i+1}^{td})}{w_1^{'} + w_2^{'} + \varepsilon} & \text{if } i = 3 \\ \frac{w_1^{'} R_i(P_i^{out}) + w_2^{'} F_i^{td} + w_3^{'} Resize(F_{i-1}^{out})}{w_1^{'} + w_2^{'} + w_3^{'} + \varepsilon} & \text{if } i = 4,5,6 \\ \frac{w_1^{'} R_i(P_i^{out}) + w_3^{'} Resize(F_{i-1}^{out})}{w_1^{'} + w_3^{'} + \varepsilon} & \text{if } i = 7 \end{cases} \tag{4}$$

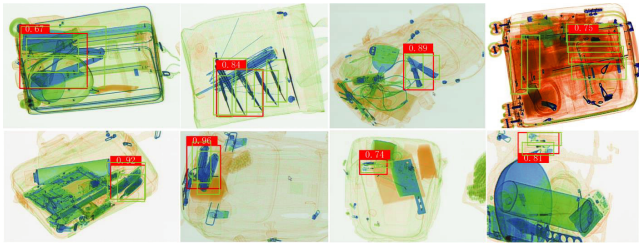


Fig. 9 Example of visualization of prohibited item detection errors. Green boxes are correctly prediction boxes, while red boxes are false positives caused by overlapping. The confidence scores outputted by detectors are also attached. The errors usually occur when a prediction box shifts slightly or dramatically to a neighboring ground-truth object, or bounds the union of several overlapping ground-truth objects.

3.4 Rep-CIoU Loss Function

The loss function used in the original cascade network for bounding box regression is Smooth L1 loss, which has some limitations in the prohibited item detection task. When the Smooth L1 loss is used to calculate the bounding box of the target detection, the losses of the four points are first calculated independently and then summed to get the final bounding box loss. While the metric in the prohibited item detection task is IoU, the Smooth L1 loss of multiple detection boxes may be the same, but the IoU may vary greatly, so the Smooth L1 loss is not applicable to the task in this paper.

In addition, overlapping prohibited items is also a problem that needs attention. As shown in Fig. 9, in the case of overlapping between multiple targets, the prediction boxes of multiple targets are regressed into one box. The reason is that the NMS algorithm filters out multiple prediction boxes because they are too close. To make each prediction box as close as possible to the ground truth while staying away from the regions of other targets, we propose the Rep-CIoU loss function of Eq. (5). Coefficients α and β act as the weights to balance the L_{CIoU} and the L_{Rep}.

$$L_{Rep-CIoU} = \alpha \cdot L_{CIoU} + \beta \cdot L_{Rep} \tag{5}$$

The L_{CIoU} loss term is expressed as Eq. (6). Where IoU is the ratio of the intersection and union of the prediction box and the ground-truth; b and b_{gt} denote the centroids of the prediction box and the ground-truth, respectively; ρ denotes the Euclidean distance; c denotes the diagonal distance of the minimum outer rectangle of the prediction box and the ground-truth; λ is a positive trade-off parameter, $\lambda = \frac{v}{(1-IoU)+v}$; v denotes the constraint on the geometric relationship of the prediction box, $v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2$, w, h, w^{gt}, h^{gt} represent the height and width of the prediction box and the height and width of the ground-truth, respectively.

The L_{Rep} loss term is expressed as Eq. (7). Where $Smooth_{ln}$ is a commonly used regression loss function, and its expression is Eq. (8); B^{P_i} and B^{P_j} denote the prediction box for the initial detection box P_i and P_j regressions; $\mathbb{1}$ is an identity function; ε is a small constant set to prevent the

divisor from being zero.

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \lambda v \tag{6}$$

$$L_{Rep} = \frac{\sum_{i \neq j} Smooth_{ln}(IoU(B^{P_i}, B^{P_j}))}{\sum_{i \neq j} \mathbb{1}[IoU(B^{P_i}, B^{P_j}) > 0] + \varepsilon} \tag{7}$$

$$Smooth_{ln} = \begin{cases} -ln(1-x) & x \leq 0 \\ x & x > 0 \end{cases} \tag{8}$$

Rep-CIoU loss function considers not only the IoU between multiple prediction boxes but also the centroid distance between prediction box and ground-truth. The L_{Rep} loss term in Rep-CIoU represents the loss value generated between a prediction box and a prediction box that is adjacent and not the same target. Its purpose is to exclude other detection boxes with different targets, making the model more robust to overlapping items. It can be found from Eq. (7) that when the IoU distance between the target prediction box P_i and other surrounding prediction boxes P_j is larger, the generated loss is also larger. Therefore, the L_{Rep} loss term can effectively prevent multiple prediction boxes from being filtered out by the NMS algorithm because they are too close to each other, and thus reduce the missed detection due to overlapping.

4. Experiments

4.1 Dataset and Experimental Setting

Dataset: The dataset includes ten types of prohibited items: Knife, Scissors, Lighter, Zippooil, Pressure, Slingshot, Handcuffs, Nailpolish, Powerbank, and Firecrackers. The visual presentation of the dataset is shown in Fig. 1(a). Figure 1(b) shows the images of five (Firecrackers, Handcuffs, Nailpolish, Slingshot, Zippooil) of the ten types of prohibited items after X-ray irradiation alone. Each of these five categories contains 200 images. The dataset comprises a total of 6,400 images, with 5,400 images of entire packages in X-ray and 1,000 images of individual prohibited items in X-ray. For the 5,400 images, the training set accounts for two-thirds and the test set accounts for one-third.

Experimental environment: The experimental environment in this paper is shown in Table 1. To control the experimental variables, we used a 32-group ResNeXt-101 (32x4d) network as the backbone network with a pre-trained model. We visualize and analyze the aspect ratio of the ground-truth of the training set images as shown in Fig. 10, so it is appropriate to set the *Anchor_Ratio* parameters in the RPN network to [0.4, 0.6, 0.8, 1.0, 2.0, 3.0].

Evaluation metrics: The mAP (mean Average Precision) is commonly used to evaluate the performance of target detection algorithms. The AP (Average Precision) is used to measure the accuracy of a certain category. The AP of all categories is averaged as mAP, and the expression is Eq. (9). Where N is the number of categories, AP_c is the AP of category c.

Table 1 Experimental environment parameters.

Name	Environment parameters
System	Linux 4.4.0-130-generic x86_64
GPU	TeslaV100-SXM2
RAM	16GB
Framework	Pytorch1.3

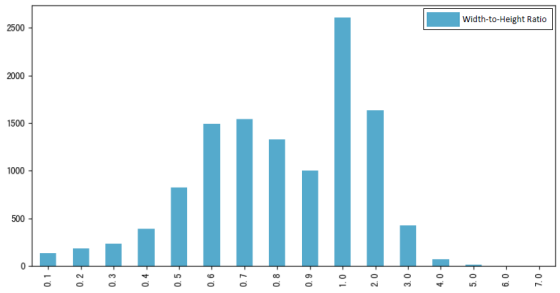


Fig. 10 Statistics on the number of ground-truth aspect ratios in the dataset. The network parameters are adjusted based on the statistics to make the network more suitable for the dataset in this paper.

Table 2 Experimental results of comparing different weighting coefficients in Rep-CIoU loss function.

α	β	mAP(%)
0.3	0.7	81.9
0.4	0.6	81.9
0.5	0.5	82.3
0.6	0.4	82.6
0.7	0.3	82.2

$$mAP = \frac{1}{N} \cdot \Sigma AP_c \tag{9}$$

4.2 Ablation Experiments of Our Proposed Method

Before the ablation experiments, we perform parametric experiments of the Rep-CIoU loss function. To verify the best performance of the Rep-CIoU loss, coefficients α and β act as the weights to balance the L_{CIoU} and the L_{Rep}. The parametric experiments are based on the original Cascade R-CNN algorithm combined with the proposed Rep-CIoU loss function to perform comparison experiments with different weighting coefficients. Table 2 shows our results with different settings of α and β. It can be concluded from Table 2 that different weighting coefficients have different effects on the algorithm accuracy. Empirically, α=0.6, β=0.4 yields the best performance.

Next, to illustrate the impact of our method on detection performance, we set up an ablation experiment with **the original Cascade R-CNN** [10] which employs the FPN and Smooth L1 loss function as the baseline. The evaluation metric is mAP, and the results of the ablation experiments are shown in Table 3. It can be seen that our proposed Poisson blending combined with the Canny edge operator method, the Re-BiFPN feature fusion method, and the Rep-CIoU loss function improved 1.5, 1.6, and 0.8 percent, respectively,

Table 3 Comparisons of the AP and mAP when adding baseline with the Poisson blending, Re-BiFPN, and Rep-CIoU.

Method	Knife	Scissors	Lighter	Zippooil	Pressure	Slingshot	Handcuffs	Nailpolish	Powerbank	Firecrackers	mAP(%)
Baseline	73.2	75.4	82.2	73.1	87.4	71.9	87.5	94.9	82.3	90.1	81.8
Baseline w/Poisson	75.8	74.4	82.4	81.4	88.5	75.2	87.6	100	81.9	86.1	83.3
Baseline w/Re-BiFPN	75.8	74.1	84.1	78.7	88.4	72.3	87.3	100	84.3	89.2	83.4
Baseline w/Rep-CIoU	75.2	75.2	82.5	72.3	88.4	74.0	87.3	96.1	85.5	89.2	82.6
Ours	78.8	78.4	84.7	78.2	89.4	81.0	87.3	100	87.3	90.9	85.6

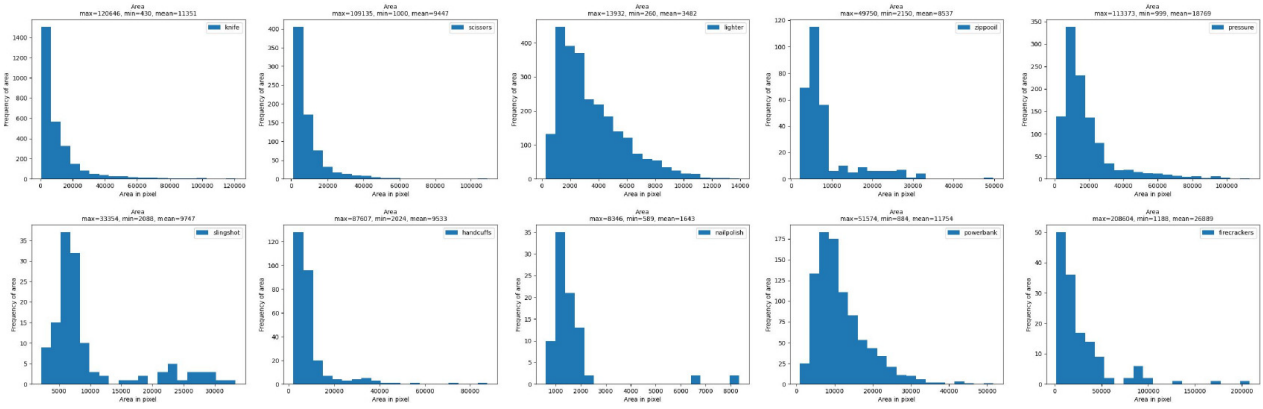


Fig. 11 Histogram of scale distribution for each category. In the subplots, the horizontal axis represents the pixel area of prohibited items, while the vertical axis represents the frequency. The headers "max", "min", and "mean" respectively indicate the maximum, minimum, and average values of prohibited item pixel areas within each category. Please note the scales of the horizontal and vertical axes in each subplot.

which seems the summation 3.9 percent could be improved theoretically. Table 3 also shows the AP for each category in the ablation experiments. It can be found that the AP of some prohibited items in each ablation experiment has improved, indicating that our method can effectively improve the accuracy of prohibited items.

4.3 Comparison Experiments between Our Method and the Baseline

Comparison of detection accuracy: Our method is the baseline combination with the Poisson blending method, the Re-BiFPN feature fusion method, and the Rep-CIoU loss function. As can be seen from Table 3, the AP has increased in most categories. The APs of Nailpolish and Firecrackers reach above 90%, and the APs of Lighter, Pressure, Slingshot, Handcuffs, and Powerbank also reach above 80%. Although the AP of Knife, Scissors, and Zippooil does not reach 80%, it is still a good improvement compared to the baseline. Compared with the baseline, the mAP of our method gets 85.6%, which is an improvement of 3.8%. Slingshot, Knife, Zippooil, and Nailpolish have the most noticeable improvement, with 9.1%, 5.6%, 5.1%, and 5.1%, respectively. Our method improves the mAP by 3.8 percent compared to the baseline. Although the theoretical improvement, 3.9 percent, is not reached, 3.8 percent can be considered as a reasonable good improvement.

In the comparison between the Baseline and Baseline with Poisson, we note that the accuracy has improved for all categories that had an increase in sample counts, except for Firecrackers. As depicted in Fig. 11, the volume of Firecrackers is larger than that of other categories, which might be related to its decrease in accuracy. For the categories Knife, Lighter, and Pressure, where the sample counts remained unchanged, their performance has also shown improvement. The enhanced accuracy for these categories could be attributed to their ample number of samples, and the improved accuracy in other categories likely reduces the risk of misclassification by the model.

In the comparison between Baseline and Baseline with Re-BiFPN, referencing Table 3 and Fig. 11, we note improvements in accuracy for several categories, including Knife, Lighter, Zippooil, Pressure, Slingshot, Nailpolish, and Powerbank. However, there was no observed improvement in accuracy for the three other categories: Handcuffs, Firecrackers, and Scissors. Drawing from the insights provided by Fig. 6(a) and Fig. 11, the decrease in accuracy for Handcuffs and Firecrackers could be attributed to their limited sample counts and a broad range of scales. The efficacy of the Re-BiFPN method might be hindered by this factor,

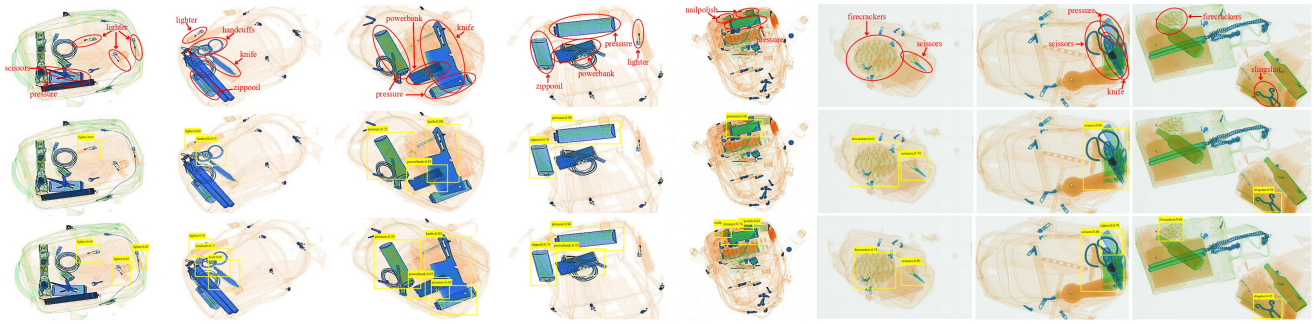


Fig. 12 Comparison of detection effect between our method and the baseline. The first row shows the input images. For clarity, we artificially highlight the prohibited items in red in the input image. The second row shows the detection effect of the baseline. The third row shows the detection effect of our method.

Table 4 Comparisons of the AP and mAP with other mainstream methods.

Method	Knife	Scissors	Lighter	Zippooil	Pressure	Slingshot	Handcuffs	Nailpolish	Powerbank	Firecrackers	mAP(%)
Mery et al. [7]	65.3	71.8	76.5	61.5	74.8	72.4	75.5	80.0	78.6	74.6	73.1
Zhang et al. [18]	71.7	66.8	78.5	68.5	83.2	77.5	86.5	80.0	83.4	86.3	78.2
Wang et al. [8]	69.4	73.5	81.3	72.6	82.5	80.5	84.7	95.0	82.5	85.2	80.7
Wang et al. [34]	71.7	75.4	78.5	72.6	83.2	78.4	86.5	90.0	87.3	85.2	80.9
Miao et al. [9]	75.4	73.5	77.9	74.5	85.5	79.6	82.5	95.0	82.5	84.8	81.1
Ours	78.8	78.4	84.7	78.2	89.4	81.0	87.3	100	87.3	90.9	85.6

given its need for a significant volume of data to adeptly learn a variety of multi-scale features. The decrease in accuracy for Scissors might be due to complex occlusions between the Scissors and the background and overlapping texture details, as illustrated in Fig. 1(a) for Scissors.

Comparison of detection effect: Figure 12 shows the detection effect of our method and the baseline.

In the image, the closer the yellow box is to the prohibited item, the better the algorithm is at locating the prohibited item. Under the condition that the labels are correct, the scores of the labels are positively correlated with the classification ability of the algorithm. Combining Table 3 and Fig. 12, we can find that the baseline (the original Cascade R-CNN) has missed detections for categories Knife, Lighter, Zippooil, Nailpolish, and Powerbank, and however our method can detect all these missed prohibited items. Moreover, even for the categories of Firecrackers, Handcuffs, Pressure, and Scissors that can be detected by the baseline, our method's localization and classification effects are better than the baseline.

4.4 Comparison Experiments between Our Method and Other Mainstream Methods

In the comparative experiments, we benchmarked against SOTA (State-of-the-Art) methods from various domains. The work by Miao [9] represents the SOTA method in the field of foreground-background separation techniques. Similarly, the method proposed in [34] stands as the SOTA method among single-stage approaches, while Wang's method [8] is recognized as the SOTA method among two-stage approaches. Zhang et al.'s technique [18] has demonstrated noteworthy performance in prohibited item detection tasks. Additionally, the method introduced by Mery [7] is a prominent representative within the domain of machine learning-based methods. These benchmarks enable us to conduct comprehensive comparative analyses to showcase the efficacy of our proposed method.

Comparison of detection accuracy: As can be seen from Table 4, the detection accuracy of our method is respectively 12.5%, 7.4%, 4.9%, 4.7%, and 4.5% higher than that of the five control groups. Among the five comparison methods, the accuracy of the algorithm proposed by Mery et al. [7] based on machine learning is only 73.1%. In contrast, the detection accuracy of our method in this paper reaches 85.6%.

Comparison of detection effect: Figure 13 shows the detection effect of our method and other comparison algorithms.

As shown in Fig. 13, the method proposed by Mery et al. [7] shows a severe wrong detection for Lighter. Moreover, Handcuffs, Nailpolish, Scissors, and Slingshot, have missed detection. Although Pressure and Firecrackers can classify correctly, their prediction boxes do not fully circle the prohibited items. The method proposed by Zhang et al. [18] also has some wrong and missed detections for Lighter, Handcuffs, Pressure, and Firecrackers. The method proposed by Wang et al. [8] has good detection for Lighter and Scissors. But Handcuffs and Pressure both have some wrong detection. Knife and Firecrackers need to be located

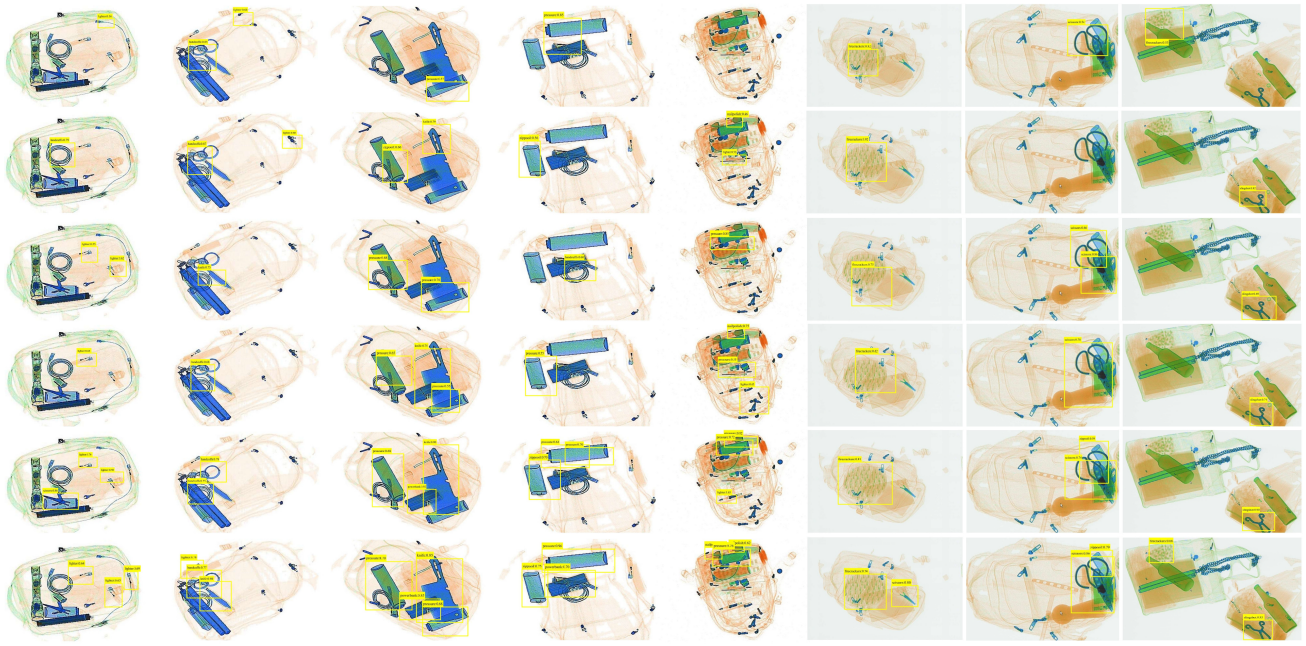


Fig. 13 Comparison of detection effect between our method and other mainstream algorithms. The input image is shown in the first row of Fig. 12. The first to fifth rows represent the detection effect of the five control groups in Table 4. The last row shows the detection effect of our method.

precisely. The method proposed by Wang et al. [34] is a representative one-stage target detection algorithm. However, Knife, Lighter, and Pressure have some detected that need to be corrected. The method proposed by Miao et al. [9] has better classification accuracy for Lighter, Scissors, Pressure, Zippooil, and Slingshot. However, it can be seen from the figure that the prediction boxes of Lighter, Pressure, and Zippooil do not accurately locate the location of the prohibited items.

The last row shows the detection effect of our method. It performs better for Lighter, Handcuffs, and Scissors, which are more prone to wrong and missed detection. In addition, our method is more accurate in locating Pressure, Firecrackers, Zippooil, and Pressure. In summary, it can be seen from Table 4 and Fig. 13 that our method has better localization precision and higher classification accuracy in this work.

5. Conclusions

In this paper, we discussed three challenges faced in prohibited item detection within X-ray security inspection images: (a) the imbalance distribution of categories, (b) diversity of prohibited item scales, and (c) overlap between items.

For (a), we proposed to leverage the Poisson blending algorithm with the Canny edge operator approach to increase the diversity and complexity of the samples. For (b), we proposed the Re-BiFPN feature fusion method, which consists of a CA-ASPP module and a recursive connection. The CA-ASPP module extracts the location information from the multi-scale feature maps. The recursive connection feeds the multi-scale feature maps processed by the CA-ASPP module to the bottom-up backbone layer. For (c), a Rep-CIoU loss function is designed to address the overlapping problem in X-ray images.

In the ablation experiments, our proposed Poisson blending combined with the Canny edge operator method, the Re-BiFPN feature fusion method, and the Rep-CIoU loss function improved 1.5, 1.6, and 0.8 percent, respectively. Comparison experiments show that our method can successfully identify ten kinds of prohibited items such as Knife, Scissors, etc., and achieved 83.4% of mAP, which is superior to the baseline (the original Cascade R-CNN) and other mainstream methods.

In our future work, we are going to further increase the types of prohibited items by adding training samples to meet the needs of different customs scenarios such as those of airports, delivery services, and subways. Additionally, we are going to further investigate and establish a consistent and objective benchmark for evaluating human visual inspection. Eventually we are going to develop a security check assistance system and deploy the model in the system to assist security staff. Such a security check assistance system can effectively reduce the labor cost and improve the quality of security inspection services.

Acknowledgments

We would like to thank iFLYTEK CO.LTD for providing us the X-ray security inspection images. This work was partially supported by JST SPRING, Grant Number JPMJSP2111.

References

[1] F. Thorsten, S. Uwe, and R. Stefan, "Object detection in multi-view X-ray images," Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium, 2012.
[2] D. Mery, E. Svec, and M. Arias, "Object recognition in baggage inspection using adaptive sparse representations of X-ray images," Image and Video Technology, pp.709–720, 2015.
[3] J. Ding, S. Chen, and G. Lu, "X-ray security inspection method using active vision based on Q-learning algorithm," Journal of Computer Applications, vol.38, no.12, pp.3414–3418, 2018.
[4] D. Mery, E. Svec, M. Arias, V. Riffo, J.M. Saavedra, and S. Banerjee, "Modern computer vision techniques for X-ray testing in baggage inspection," IEEE Trans. Syst. Man Cybern., Syst., vol.47, no.4, pp.682–692, 2016.
[5] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded prohibited item detection: An X-ray security inspection benchmark and de-occlusion attention module," Proc. 28th ACM International Conference on Multimedia, pp.138–146, 2020.
[6] M. Baştan, "Multi-view object detection in dual-energy X-ray images," Machine Vision and Applications, vol.26, pp.1045–1060, 2015.
[7] D. Mery, G. Mondragon, V. Riffo, and I. Zuccar, "Detection of regular objects in baggage using multiple X-ray views," Insight-Non-Destructive Testing and Condition Monitoring, vol.55, no.1, pp.16–20, 2013.
[8] B. Wang, L. Zhang, L. Wen, X. Liu, and Y. Wu, "Towards real-world prohibited item detection: A large-scale X-ray benchmark," Proc. IEEE/CVF International Conference on Computer Vision, 2021.
[9] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. Ye, "SIXray: A large-scale security inspection X-ray benchmark for prohibited item discovery in overlapping images," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
[10] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[11] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[12] F. Shao, J. Liu, P. Wu, Z. Yang, and Z. Wu, "Exploiting foreground and background separation for prohibited item detection in overlapping X-Ray images," Pattern Recognition, vol.122, 108261, 2022.
[13] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded prohibited items detection: An X-ray security inspection benchmark and De-occlusion attention module," Proc. 28th ACM International Conference on Multimedia, pp.138–146, 2020.
[14] T. Hassan, H. Khan, and S. Akcay, "Deep CMST framework for the autonomous recognition of heavily occluded and cluttered baggage items from multivendor security radiographs," arXiv preprint arXiv:1912.04251, 2019.
[15] S. Akcay, M.E. Kundegorski, M. Devereux, and T.P. Breckon, "Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery," 2016 IEEE International Conference on Image Processing (ICIP), 2016.
[16] M. Roomi and M. Rajashankarii, "Detection of concealed weapons in X-ray images using fuzzy K-NN," International Journal of Computer Science, vol.2, no.2, pp.187–196, 2012.
[17] S. Akcay, M. Kundegorski, C. Willcocks, and T. Breckon, "Using deep convolutional neural network architectures for object classification and detection within X-ray baggage security imagery," IEEE Trans. Inf. Forensics Security, vol.13, no.9, pp.2203–2215, 2018.
[18] Y. Zhang, Z. Su, H. Zhang, and J. Yang, "Multi-scale prohibited item detection in X-ray security image," Journal of Signal Processing, vol.36, no.7, pp.1096–1106, 2020.
[19] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[20] M. Tan, R. Pang, and Q. Le, "EfficientDet: Scalable and efficient object detection," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[21] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
[22] K. Chen, Z. Zhu, X. Deng, C. Ma, and H. Wang, "Deep learning for multi-scale object detection: A survey," Journal of Software, vol.32, no.4, pp.1201–1227, 2021.
[23] X. Wang, S. Zhang, Z. Yu, L. Feng, and W. Zhang, "Scale-equalizing pyramid convolution for object detection," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[24] K. Min, H. Lee, and S. Lee, "Attentional feature pyramid network for small object detection," Neural Networks, vol.155, pp.439–450, 2022.
[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[26] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," Proc. 24th ACM International Conference on Multimedia, pp.516–520, 2016.
[27] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[28] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
[29] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," Proc. AAAI Conference on Artificial Intelligence, vol.34, no.7, pp.12993–13000, 2020.
[30] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," arXiv preprint arXiv:2005.03572, 2020.
[31] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[32] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," ACM SIGGRAPH 2003 Papers, pp.313–318, 2003.
[33] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[34] Z. Wang, H. Zhang, Z. Lin, X. Tan, and B. Zhou, "Prohibited items detection in baggage security based on improved YOLOv5," 2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI), 2022.

Qingqi Zhang was born in 1996. He received the B.E. from Heilongjiang University, China, in 2019 and the M.E. from Zhejiang Sci-Tech University, China, in 2022. He is currently a Ph.D. candidate in the Graduate School of East Asian Studies, Yamaguchi University, Japan. His main research interests include computer vision and pattern recognition.

Xiaoan Bao received the B.S. from Zhejiang University (China) in 1998, and M.S. from China West Normal University in 2004. He was an Associate Professor at Zhejiang Sci-Tech University, China, from 2007 to 2012. Since November 2012, he has been a Professor at Zhejiang Sci-Tech University, China. His main research interests include software engineering and computer vision, and pattern recognition.

Ren Wu received B.E. and M.E. from Hiroshima University, Japan, in 1988 and 1990, respectively, and Ph.D. from Yamaguchi University, Japan, in 2013. She was with Fujitsu Ten Ltd., West Japan Information Systems Co., Ltd. and Yamaguchi Junior College from 1991 to March 2024. Since April 2024, she has been an Associate Professor at Shunan University, Japan. Her research interest includes information processing systems, linguistic information processing and system modeling. She is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Institute of Information Processing Society of Japan (IPSJ).

Mitsuru Nakata received B.E., M.E. and Ph.D. from Fukui University, Japan, in 1992, 1994 and 1998, respectively. He was a Lecturer from 1998 to 2004 and an Associate Professor from 2004 to 2014 both at Yamaguchi University, Japan. Since October 2014, he has been a Professor at Yamaguchi University. His research interest includes database system, text processing and program net theory and information education. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Institute of Information Processing Society of Japan (IPSJ) and the Institute of Electrical and Electronics Engineers (IEEE).

Qi-Wei Ge received B.E. from Fudan University, China, in 1983, M.E. and Ph.D. from Hiroshima University, Japan, in 1987 and 1991, respectively. He was with Fujitsu Ten Limited from 1991 to 1993. He was an Associate Professor at Yamaguchi University, Japan, from 1993 to 2004. Since April 2004, he has been a Professor at Yamaguchi University, Japan. He is currently a Trustee at Yamaguchi University, Japan. His research interest includes Petri nets, program net theory and combinatorics. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Institute of Information Processing Society of Japan (IPSJ) and the Institute of Electrical and Electronics Engineers (IEEE).