electronics

MDPI

Article

Material-Aware Path Aggregation Network and Shape Decoupled SIoU for X-ray Contraband Detection

Nan Xiang [1], Zehao Gong [1], Yi Xu [1] and Lili Xiong [2,*]

[1] Liangjiang International College, Chongqing University of Technology, Chongqing 400054, China
[2] Chongqing Academy of Science and Technology, Chongqing 401331, China
* Correspondence: sealilyxiong@163.com

Abstract: X-ray contraband detection plays an important role in the field of public safety. To solve the multi-scale and obscuration problem in X-ray contraband detection, we propose a material-aware path aggregation network to detect and classify contraband in X-ray baggage images. Based on YoloX, our network integrates two new modules: multi-scale smoothed atrous convolution (SCA) and material-aware coordinate attention modules (MCA). In SAC, an improved receptive field-enhanced network structure is proposed by combining smoothed atrous convolution, using separate shared convolution, with a parallel branching structure, which allows for the acquisition of multi-scale receptive fields while reducing grid effects. In the MCA, we incorporate a spatial coordinate separation material perception module with a coordinated attention mechanism. A material perception module can extract the material information features in X and Y dimensions, respectively, which alleviates the obscuring problem by focusing on the distinctive material characteristics. Finally, we design the shape-decoupled SIoU loss function (SD-SIoU) for the shape characteristics of the X-ray contraband. The category decoupling module and the long–short side decoupling module are integrated to the shape loss. It can effectively balance the effect of the long–short side. We evaluate our approach on the public X-ray contraband SIXray and OPIXray datasets, and the results show that our approach is competitive with other X-ray baggage inspection approaches.

Keywords: X-ray images; contraband detection; atrous convolution; attention mechanism; regression loss function

Citation: Xiang, N.; Gong, Z.; Xu, Y.; Xiong, L. Material-Aware Path Aggregation Network and Shape Decoupled SIoU for X-ray Contraband Detection. *Electronics* **2023**, *12*, 1179. https://doi.org/10.3390/electronics12051179

Academic Editor: Eva Cernadas

Received: 21 January 2023
Revised: 17 February 2023
Accepted: 24 February 2023
Published: 28 February 2023

1. Introduction

With the development of the transportation industry, transportation security has become a key area of concern, where contraband detection is an important measure to maintain public safety and transportation security. However, the current excessive reliance on the experience and energy of security personnel has decreased the accuracy of manual reviews, and the accuracy rate of contraband detection by security personnel is generally between 80% and 90% [1]. Therefore, automatically searching for prohibited items in passenger packages from X-ray images is essential for reducing labor costs and improving efficiency and reliability.

Through the analysis of the dual-energy X-ray scanning contraband dataset and operation of related experiments, it is found that they compared with the photographic (optical) object detection dataset, MS-COCO [2] (Microsoft Common Object in Context), and the dataset PASCAL VOC [3]. In the past few years, artificial intelligence technology based on the neural network has been applied to X-ray contraband detection [4–6]. However, these algorithms have not yielded satisfactory achievements in contraband detection. Contraband security screening remains an open challenge for several key reasons [7]:

1. Multi-scale detection in X-ray datasets: Due to the scanning angle of the dual-energy X-ray scanner and the physical characteristics of the contraband, there is a seriously uneven scale, which includes an uneven scale between the different categories, an

uneven scale between the same categories, and an uneven scale between the long–short sides, rendering it difficult to detect the contraband.

2. Extreme clutter and occlusion: Pieces of information obscure each other because of the penetrating nature of the X-ray scanning equipment and the resulting overlap between the deep and shallow high-density image. This has a negative impact on the accuracy of X-ray contraband detection.

To solve the above problems, this paper proposes a material-aware path aggregation network for X-ray object detection and shape-decoupled SIoU (SD-SIoU), which can not only detect items of contraband in common but also detect difficult samples in extreme cases, such as small objects and obscured items. Our model takes the YoloX [8] object detection network as the baseline and modifies its neck part for the differences between the X-ray images and the natural images in the OPIXray [9] dataset. Figure 1 shows the images of the dataset with the above problem.

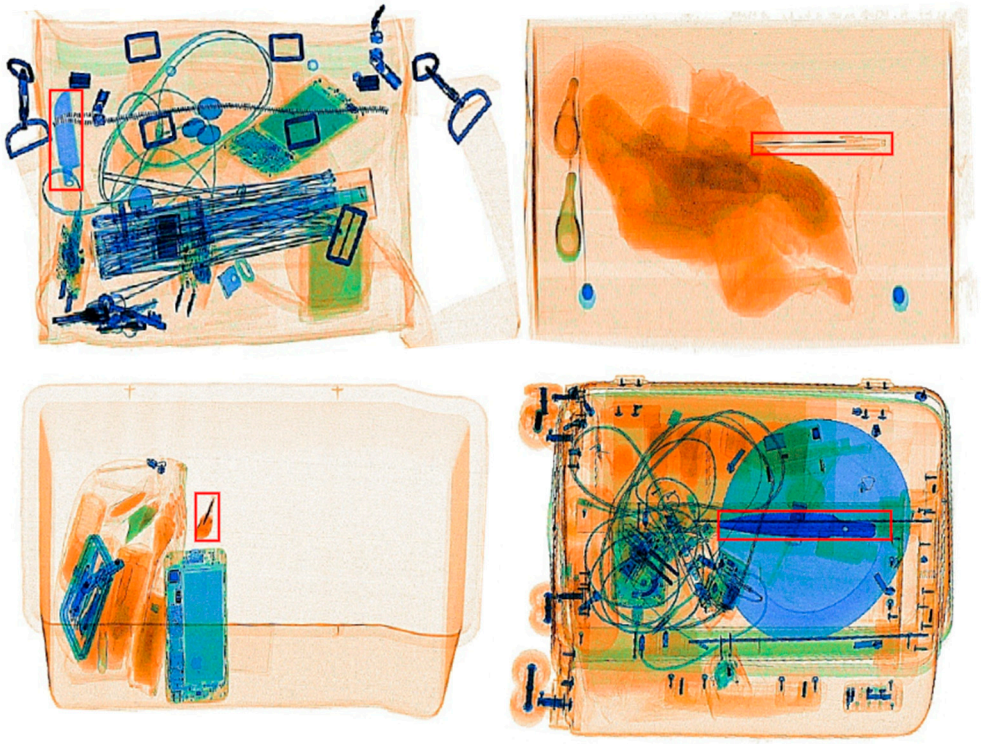


Figure 1. Problem description in X-ray contraband dataset. The first three images show the scale difference problem caused by different views of the same type of contraband and its uneven aspect ratio, and the last image shows the complex occlusion and clutter problem.

Our main contributions are listed below:

1. Constructing a novel material-aware path aggregation network, which includes a smoothed atrous convolution module (SAC) and material-aware coordinate attention mechanism (MCA). The SAC is to handle the multi-scale problem by combining smoothed atrous convolution using separate shared convolutions with a parallel branching structure. The SAC effectively mitigates the grid effect caused by the atrous convolution, while improving the model's multi-scale detection capability. The MCA is designed to address the clutter and occlusion problem by incorporating a spatial coordinate separation material perception module with a coordinate attention mechanism. The MCA mitigates contraband obstruction by focusing deeply on the contraband material information.
2. A new shape-decoupled SIoU (SD-SIoU), based on the SIoU, is constructed for the uneven aspect ratio problem. First, we optimize the normalized penalty factor; a cen-

trosymmetric normalization function is constructed. Then, we decouple the predicted bounding box long–short side length information to construct a long–short-shape loss branch. Finally, we introduce the category long–short side coefficient, which is determined by category prior knowledge of the contraband datasets. The category long–short coefficient is embedded in the long–short-shape loss branch to handle the uneven aspect ratio by utilizing the category prior knowledge.

3. We evaluate our module on the OPIXray [9] and SIXray [10] datasets, then compare it to recent high-performing object detection networks and contraband detection networks. The experimental results confirm the superiority of our model over other contraband detection models.

2. Relate Work

X-ray security inspection task. Compared to the traditional photographic imagery generated by light reflection, an X-ray image is based on X-ray properties (penetrating, fluorescent and photographic effects). In X-ray images, the brightness and color of the pictures represent the density and material of the detected items, respectively. Therefore, objects scanned by X-ray lose their texture and original color information.

Traditional feature detection methods. X-ray contraband detection belongs to the category of object detection, and the early object detection feature extractors were mostly designed manually and purposefully. Turcsany et al. [11] used a Support Vector Machine (SVM) and SURF features (Speeded-UP Robust Features) to build a visual bag-of-words; Zhang et al. [12] extracted potential features of the image, such as the edges and color, by traditional image processing methods, and obtained a good detection performance improvement.

Deep learning detection methods. Deep learning comprises multiple layers of neural networks that outperform traditional machine learning algorithms. Akcay [13] et al. first introduced deep learning to luggage classification detection of X-ray images using transfer learning. Li et al. [14] combined a semantic segmentation network with Mask R-CNN [15] into a two-stage CNN model, using the semantic segmentation network as Mask R-CNN soft-attention coding to improve the performance degradation caused by overlapping objects in X-ray images. Zhang et al. [16] used an XMC R-CNN model, consisting of a material classification algorithm and an organic-inorganic separation algorithm, for object detection to mitigate the accuracy degradation caused by the occlusion problem effectively.

Multi-scale problem in contrabands detection. Few research studies focus on X-ray baggage threat detection in complex scenarios, including multi-scale detection. Wang et al. [17] utilized a dense attention module to contribute to SDANet, and Cascade Mask RCNN is used as the baseline for the extracted multi-scale features. Tao et al. [18] utilized bidirectional propagation to filter out the impact of the noisy region in the key part by constructing multi-scale features links. Chunjie et al. [19] proposed EAOD-Net, utilizing the learnable Gabor convolution and deformable convolution. ResNeXt is also used to improve the representative ability of multi-scale features. Nguyen et al. [20] used a task-specific deep feature extractor to reduce the multi-scale X-ray images to the same aspect ratio in the same size. This can enable a more efficient deep-detection pipeline. Chunjie et al. [21] constructed a global context feature extraction (GCFE) module and learnable Gabor convolution layer for the high-level and low-level features, which facilitates the detection of bands of different sizes while suppressing background noise.

Obscuration problem in contrabands detection. The obscuration problem has also been widely studied by many scholars. Gas et al. [22] explored the ability of the traditional CNN model to adapt different properties of the scanner and evaluated the prohibited items predicted result on the Dbf3 and SIXray datasets. Hassan et al. [23] obtained dual tensors with improved contour information in X-ray baggage images by levering the intensity transit transitions in low- and high-energy scans. Those contour features were then put into an edge suppression model to filter the noise information to a normal level. Li et al. [24] proposed a method based on GANs with a generator architecture with Res2Net for the natural occurrence problem. Hassan et al. [25] proposed a tensor pooling strategy to

decompose the scans across various scales and then fuse them via a single multi-scale tensor to obtain more salient contour maps for boosting a framework's capacity for handling the overlap problem. Wei et al. [9]. proposed the de-occlusion module (DOAM), which combines the edge and material information of the contraband to refine the feature map, which enhances the detection performance.

However, edge information contains too many irrelevant gradients [26]. Therefore, it has a limited improvement in the model localization and classification; this leads to poor discrimination by the detection model in the case of occlusion and a multi-scale task. In addition, the above model does not take into account the effect of a severely unbalanced aspect ratio on the model predictions, which prevents the model from using the contraband shape information distribution to improve the model's prediction performance.

3. Method

The anchor-free detection method is able to learn multi-scale features better than the anchor-based method [27]. Therefore, the YoloX model using the anchor free detection method is chosen as the baseline model in this paper. A new shape-decoupled SIoU loss is also designed for YoloX's unique decoupling.

The block diagram of the proposed framework is depicted in Figure 2. The input origin image is fed into the CSP-DarkNet53 [28] backbone for multi-scale feature extraction. The extracted multi-scale features are separately fed into the material-aware coordinate attention mechanism (MCA) for recalibration. In the MCA, the material information related to the contraband can be extracted and integrated more accurately by utilizing a spatial coordinate separation material perception module. Afterward, these features containing the aggregated material information are then fed into an improved path aggregation network (PAN) [29], which is embedded in the multi-scale smoothed atrous convolution module (SAC), with the SAC levering the ability of the smoothed atrous convolution to increase the field of perception for further extraction and fusion of multi-scale object information. Finally, in the training stage, the contraband prediction results are output by the decoupling head. SD-SIoU is used in the bounding box loss calculation, which decouples the shape loss of the prediction box into the long-side and short-side shape loss. The specific details will be described in the following sections.

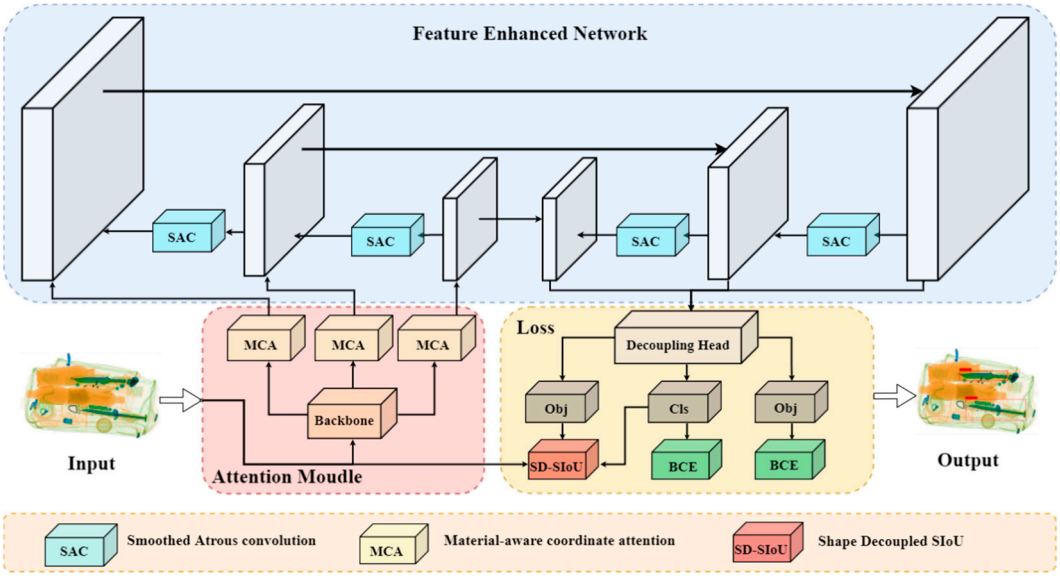


Figure 2. Overall architecture diagram. The network backbone for the feature extraction is CSP-Darknet53, which is the same in the detection network. The feature-enhanced network is a path aggregation network with a SAC module.

3.1. Material-Aware Path Aggregation Network

To further address the problem of multi-scale detection and occlusion in contraband images, a Material-aware Path Aggregation network is proposed, which consists of multi-scale smoothing atrous convolution (SAC) and a material-aware coordinate attention mechanism module (MCA).

3.1.1. Multi-Scale Smoothing Atrous Convolution (SAC)

Compared to the traditional convolutional method, atrous convolution increases the receptive field of the convolution kernel while keeping the number of parameters unchanged [30]. However, atrous convolution faces a serious grid effect, weakening the proximate connections while gaining long-distance dependence. To address this problem, inspired by the smoothed atrous convolution [31], a multi-scale parallel smoothed atrous convolution structure is designed, which is shown in Figure 3.

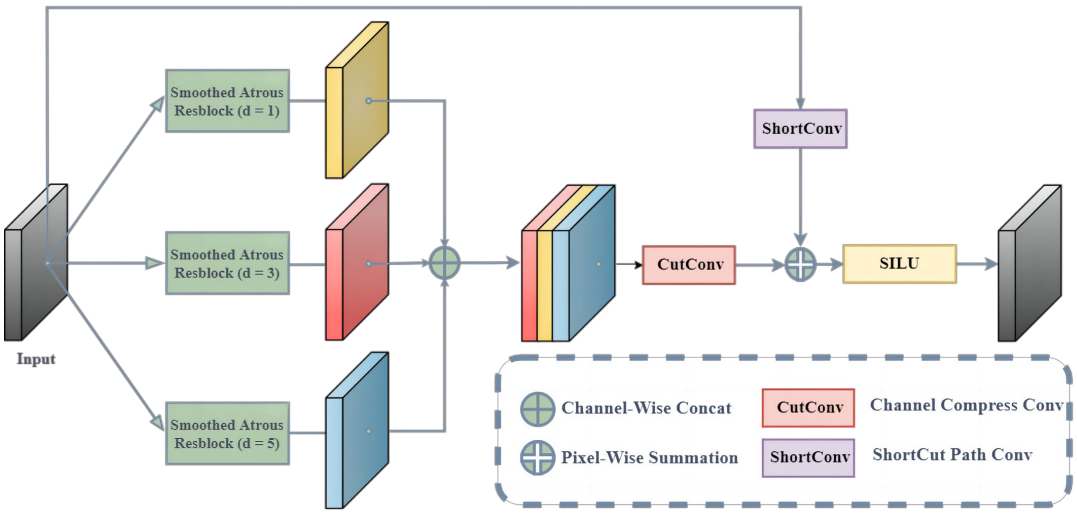


Figure 3. Smoothed atrous convolution (SAC) structure diagram.

As shown above, to limit the impact of the grid effect, this paper constructs parallel atrous convolution branches; each branch uses a different expansion rate to minimize the grid effect. Figure 4 shows the visualization of the atrous convolution grid effect rendering.

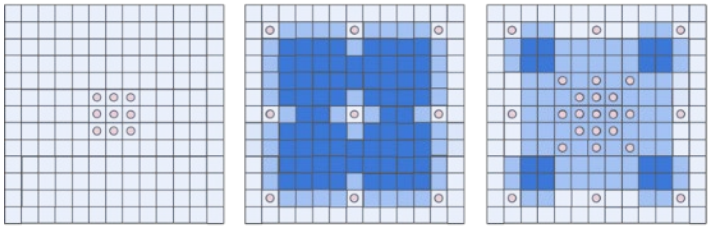


Figure 4. Visualization of receptive fields and grid effects.

A smoothed dilated residual block can effectively prevent the grid effect [31]; it addresses the gridding effect by levering separable and shared convolutions (SS), based on the idea of separable convolutions [32]. In SS convolutions, sharing means that the filters are the same and shared by all the input and output channel pairs. For both the input and output channels, the SS convolution uses only one filter to obtain all the spatial information and shares that filter over all the channels. Therefore, smoothed dilated convolutions can effectively amplify the receptive field to make this branch pay more attention to style features(e.g. edges and global colors) [33]. We therefore apply this module to our parallel multi-scale architecture. Finally, inspired by ResNet [34], the residual

information is summed with the fused information in Pixel-Wise and activated by the SiLU activation function.

Although the use of null convolution is effective in reducing the computational effort, the model itself increases some of the parameters and computational effort because of the addition of extra convolution

By using the SAC in the path aggregation network, the weight of the contraband material information can be augmented, which significantly increases the capability of the features to describe the important objects.

3.1.2. Material-Aware Coordinate Attention Mechanism (MCA)

Due to the unique physical characteristics of the X-ray scanner, the material information of the contraband is greatly diminished and is ultimately represented as color information. This means that channel information has a greater contribution to the detection of contraband in X-ray scanned images. The channel attention mechanism can learn different weights of channel dimensions, so that the information from the key channels can be utilized to a greater extent. The coordinate attention (CA) mechanism [35], as a kind of channel attention module, embeds the spatial location information into the channel attention, which means adding extra information into the channels.

However, due to the weakness of spatial information in X-ray images, the original CA attention mechanism cannot fully extract the comprehensive spatial information of images. For this problem, inspired by SRM [36], a material-aware coordinate attention mechanism is designed, and the specific structure is shown in Figure 5.

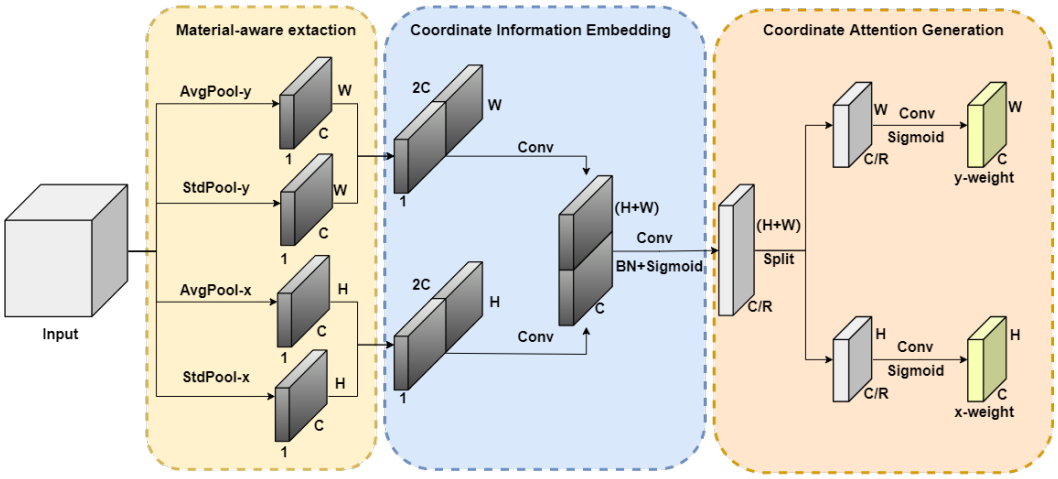


Figure 5. Material-aware coordinate attention (MCA) structure diagram.

First, the input feature maps are put into the material-aware extraction module, which is constructed by average pooling and standard pooling in the width and height directions, to obtain four feature maps, respectively. Specifically, given the input X, two special two-dimensional convolution kernels, (H,1) and (1, W), are used to encode the input data, and four different pooling methods are used to obtain the horizontal and vertical coordinate encoding information. The output of the height, h, at the c-th channel can be presented as

$$Avg_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{1}$$

$$Std_c^h(h) = \sqrt{\frac{1}{W} \sum_{0 \le i < W} \left(x_c(h, i) - Avg_c^h(h)\right)^2} \tag{2}$$

Similarly, the output of width, w, at the c-th channel can be formulated as

$$Avg_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{3}$$

$$Std_c^w(w) = \sqrt{\frac{1}{H} \sum_{0 \le j < H} \left(x_c(j, w) - Avg_c^w(w)\right)^2} \tag{4}$$

The above four branches integrate the information of two spatial dimensions, encoding the spatial information and channel information together. This serves as a summary description of the material information for each example, n, and channel, c.

After that, we enter the coordinate information embedding layer to splice and convolve the channel dimensions of the width and height feature information, which embeds the width and height information with the channel information into one feature map. Two feature maps with scales of H × 1 × C and 1 × W × C are obtained. These two directional feature maps of the width and height of the obtained global receptive field are put together according to the spatial dimension. Then, in the coordinate attention generation part, the two feature maps are fed into a convolution module with a shared convolution kernel of 1 × 1 to scale the dimension to C/r and, finally, to the sigmoid activation function and the BatchNorm operation.

3.2. Shape Decoupling SIoU (SD-SIoU)

In addition to the anchor-free detector, YoloX also introduces a decoupled head. The decoupled head decouples the classification task and the regression localization task into two separate branches for separate outputs. This enables the model to focus on the classification and localization tasks separately and improve the model performance. We further improve the decoupled localization task by introducing the SioU [37] loss function and improving it for the physical properties of the X-ray scanning object, which include the shape-decoupling module and normalized optimization algorithm

3.2.1. Revisit SIoU Loss Function

Traditional IoU losses, such as DIoU, CioU [38] and GioU [39], only consider the distance, overlap area and aspect ratio information, and do not consider the angle and ratio between the shape and the predicted bounding box and the target bounding box, resulting in a slight overlap. However, SIoU redefines the penalty matrix by considering the angle and shape. SIOU regression loss consists of four components: distance loss, IOU loss, angle loss and shape loss. The total loss is defined as:

$$L_{\text{box}} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{5}$$

The angle loss is defined as:

$$\Lambda = 1 - 2 * \sin^2\left(\arcsin(x) - \frac{\pi}{4}\right) \tag{6}$$

$$x = \frac{\max(b_{cy}^{gt}, b_{cy}) - \min(b_{cy}^{gt}, b_{cy})}{\sqrt{\left(b_{cx}^{gt} - b_{cx}\right)^2 + \left(b_{cy}^{gt} - b_{cy}\right)^2}} \tag{7}$$

The distance loss is defined as:

$$\Delta = \sum_{t=x,y} \left(1 - e^{(\Lambda - 2)\rho_t}\right) \tag{8}$$

where

$$\rho_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{\max(w, w^{gt})}\right)^2, \rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{\max(h, h^{gt})}\right)^2 \tag{9}$$

The shape loss is defined as:

$$\Omega = \sum_{t=w,h} \left(1 - e^{-\omega_t}\right)^{\theta} \tag{10}$$

where

$$\omega_w = \frac{|w - w^{gt}|}{\max(w, w_{gt})}, \omega_h = \frac{|h - h^{gt}|}{\max(h, h_{gt})} \tag{11}$$

b_{cy}^{gt} and b_{cy} represent the y coordinates of the center point for ground truth and prediction. w and h represent the width and height of the bounding box.

SIoU has been widely used in recent networks and has proven to be a key component in the implementation of advanced detectors [40–43]. However, although SIoU takes shape loss into account, it couples the long- and short-side information of the prediction bounding box together and assigns the same computational weight to them, which ignores the proportional relationship between the long and short sides. In addition, SIoU limits the shape loss to [0, 1] by dividing by the maximum of the predicted and true values, which causes asymmetry in the parameter convergence curve and convergence difficulties due to low proximity gradients.

In the following, we will reconsider the shape loss part for the above problem.

3.2.2. Shape Decoupling Module

In the X-ray contraband images, the distribution of the long side and short side is always not equal, and the aspect weight of contraband varies greatly among different categories. Giving the same weight to the long side and short side will affect the optimization of the model for the contraband shape information. Figure 6 shows the scatter plot of the OPIXray dataset consisting of information on the long side and short side of different types of contraband.

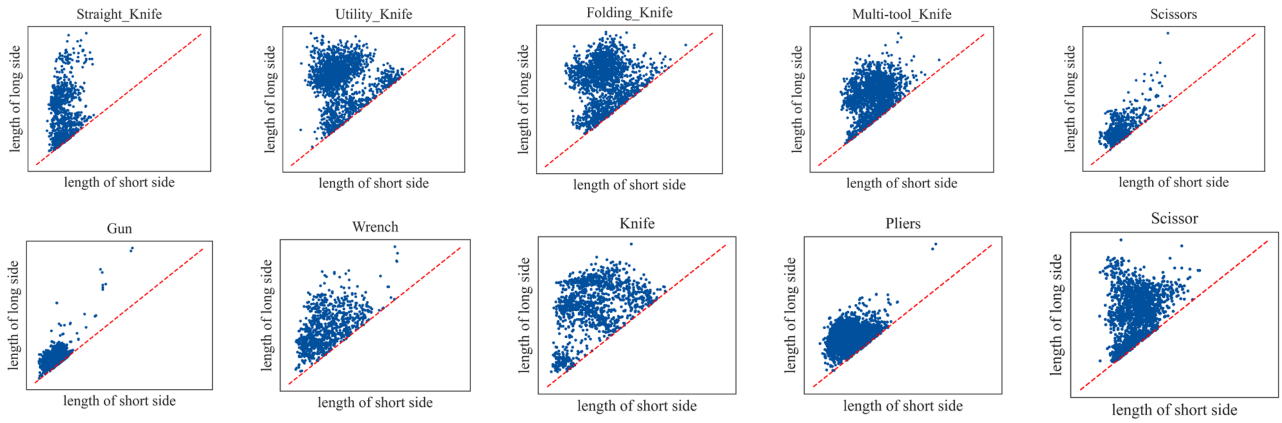


Figure 6. Comparison of length ratio of different categories under SIXray and OPIXray datasets.

As shown in Figure 6, there is a significant difference between the long–short sides of the target box. To address this problem, we designed the long–short side decoupling module and the category information embedding module, based on the special structure of the YoloX decoupling head. The detailed structure is shown in Figure 7.

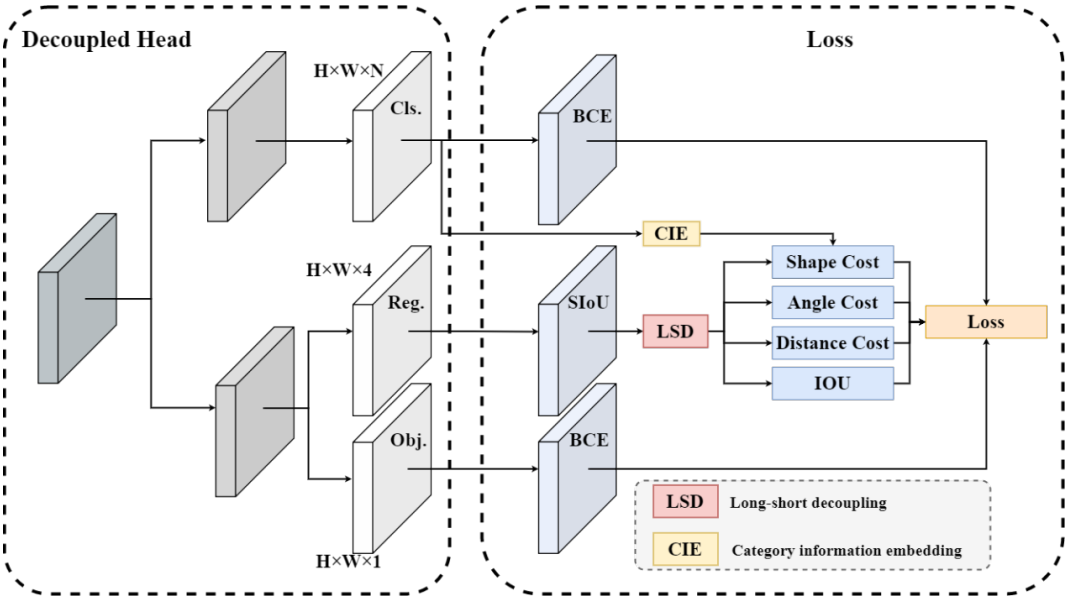


Figure 7. SD-SIoU structure diagram.

In the long–short side decoupling module, the length and width information of the input prediction bounding box is separated, and the lengths of the long side and the short side are extracted, respectively. Therefore, a new shape loss penalty factor is decoupled for the long length, l, and short length, s, as follows.

$$\omega_l = \frac{|l - l^{gt}|}{\max(l, l_{gt})}, \omega_s = \frac{|s - s^{gt}|}{\max(s, s_{gt})} \tag{12}$$

In the category information embedding module, we collect the long and short side information of the dataset by category and perform a cluster analysis to obtain the gathering point information. Finally, we construct the long–short scale matrix, $M_{n\times1}$, which can be represented as follows.

$$M = [\alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_n]_{n\times1} \tag{13}$$

where n is the number of categories, and α_i is the aspect ratio of i-th category clustered. Then, multiplying the category prediction matrix, $C_{m\times n}$, with the long–short scale matrix, $M_{n\times1}$, yields the category long–short side coefficient matrix, $A_{m\times1}$.

Then, we embed the category information into the shape loss by dividing the long-side penalty factor by the category long–short side coefficient matrix, $A_{m\times1}$. The equation is shown below.

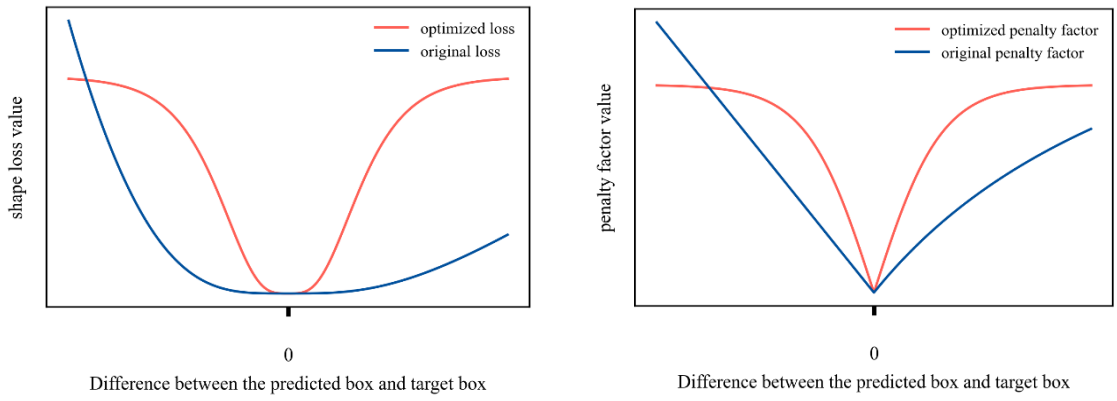
$$\omega_{l+} = \omega_l / A_{m\times1} \tag{14}$$

The above formula realizes the decoupling of the shape information and the embedding of the category information, effectively alleviating the impact of the long–short sides on detection accuracy.

3.2.3. Normalized Optimization Module

As we continue our research, we find that, in shape loss, the range of values is restricted to $\mathbb{R} \in (0, 1)$ by dividing by the maximum value of the ground truth box width and height and the predicted box width and height in Equation (7). However, this method leads to a symmetry problem. It can be seen, in Figure 8, the maximum normalization does not work consistently for the same distance gap between the target and predicted bounding box sizes in the positive and negative directions, and the optimized gradient is worse as distance between the target and prediction gets closer. Although the function has a very

fast convergence speed in the early stage of training, the convergence ability of the model decreases as the prediction results approach.



(**a**) (**b**)

Figure 8. Comparison of shape loss and normalized penalty factor before and after improvement. (**a**) Shape loss; (**b**) normalized penalty factor.

To address this problem, we designed a symmetric normalization method for the shape loss part of the SIoU. The new shape loss composition is shown below.

$$\Omega = \sum_{t=w,h} \left(1 - e^{-\varphi_t}\right)^{\theta} \tag{15}$$

where the novel penalty factor is:

$$\omega_t^+ = b \times \frac{e^{k\times|w-w^{gt}|} - e^{-k\times|w-w^{gt}|}}{e^{k\times|w-w^{gt}|} + e^{-k\times|w-w^{gt}|}} \tag{16}$$

As shown above, the improved normalization function solves the left–right asymmetry problem caused by the max function and optimizes the penalty factor regularization algorithm, so that the loss decreases more smoothly during the training process and still has a certain descent gradient in the late training period.

4. Experiment

In this section, we conduct comprehensive experiments on OPIXray and SIXray datasets to evaluate the effectiveness of our method. OPIXray and SIXray are the common datasets for X-ray contraband images.

4.1. Experiment Setting Details

This paper is implemented by a Windows 10 64-bit operating system, 12th Gen Intel Core i9-12900K@3.2 GHz CPU, 32 GB RAM, NVIDIA 3080ti GPU with CUDA Toolkit 11.4 and Torch 1.11 in Python 3.8. As the benchmark of our model, YoloX uses the most primitive parameter settings. The backbone of YoloX uses CSP-Darknet53.

All the experiments of our model and baselines are optimized by an Adam optimizer. The initial learning rate is set to 0.001, and the Cosine Annealing learning rate reduction strategy is used. The momentum and weight decay are set to 0.93 and 0, respectively. The batch size is set to 16. We evaluate the mean Average Precision (mAP) to measure the performance of all the methods. In addition, the IoU threshold measuring the accuracy of the predicted bounding box is set to 0.5.

4.2. Comparing with SOTA Detection Methods

To verify the effectiveness of the proposed methods in this paper, as shown in Tables 1 and 2, we compared the mainstream contraband detection models and object detection models in the last two years on the OPIXray and SIXray datasets, respectively. The method involved included object detection models such as Swin Transformer [44], RetinaNet [45], DetectoRS [46], Yolov5 and baseline YoloX. It also includes the most advanced contraband detection models in the last two years such as CHR [10], FBS [47], CFPA-Net [48], MCIA-FPN [49] and POD-Y [21].

Table 1. Performance comparison results using different object detection methods on the OPIXray dataset.

Model	Year	Backbone	Category					mAP
			FO	ST	SC	UT	MU	
Swin Trans [44]	2021	Swin Trans	82.14	42.77	95.75	69.60	84.84	75.04
CHR [10]	2019	Resnet-50	87.94	84.53	95.23	50.99	74.47	78.63
RetinaNet [45]	2017	Resnet-50	89.27	55.66	98.15	79.79	85.27	81.63
FBS [47]	2022	CSPDarknet53	86.38	88.29	95.45	57.99	80.62	81.75
CFPA-Net [48]	2021	Resnet-50	87.72	76.10	90.52	85.94	84.87	81.84
DetectoRS [46]	2021	Resnet-50	88.51	64.01	89.86	81.02	86.59	82.00
DOAM [9]	2020	Resnet-50	86.71	68.58	90.23	78.84	87.67	82.41
Yolov5	2021	CSPDarknet53	90.36	64.85	97.69	80.93	94.44	85.65
MCIA-FPN [49]	2022	ResNet-101	89.08	74.48	89.99	86.13	89.75	85.89
ATSS-Lacls [50]	2022	ResNet-50	92.31	72.04	96.58	80.23	91.67	86.59
Chang et al. [5]	2022	Resnet-50	90.42	75.95	91.46	84.31	91.29	86.69
YoloX [8]	2021	CSPDarknet53	91.84	77.53	97.89	89.22	92.79	89.85
LIM [18]	2021	Resnet-50-FPN	94.79	77.66	98.20	88.92	93.75	90.43
POD-Y [21]	2022	CSP-Darknet53	94.5	77.8	98.2	89.5	94.5	90.9
Ours	N/A	CSPDarkNet53	94.53	86.68	98.88	89.56	94.96	92.92

Table 2. Performance comparison results using different object detection methods on the SIXray dataset.

Model	Year	Backbone	Category					mAP
			Gun	Knife	Wrench	Pliers	Scissors	
CHR [10]	2019	Resnet50	79.22	63.77	73.77	71.55	65.55	70.77
RetinaNet [45]	2017	Resnet-50	81.16	77.27	33.24	66.87	22.61	81.50
FBS [47]	2022	CSP-DarkNet53	79.72	64.14	74.96	71.19	66.17	71.24
DetectoRS [46]	2021	Resnet-50	81.61	80.52	84.48	87.40	81.4	83.10
CFPA-Net [48]	2021	Resnet-50	86.07	86.33	72.44	87.28	75.95	81.61
DOAM [9]	2020	CSP-Darknet53	81.37	64.25	73.26	70.17	61.98	70.21
MCIA-FPN [49]	2022	Resnet101	85.75	83.75	81.50	86.79	88.34	85.23
Yolo v5	2021	CSP-Darknet53	97.36	84.60	90.00	85.56	85.20	88.55
YoloX [8]	2021	CSP-Darknet53	96.74	85.94	91.48	86.94	87.89	89.80
POD-Y [21]	2022	CSP-Darknet53	92.6	87.9	87.6	92.1	91.8	90.4
Ours	N/A	CSP-Darknet53	97.01	87.63	88.66	92.48	89.70	91.10

As Tables 1 and 2 show, the proposed model can achieve the optimal detection performance on the OPIXray and SIXray datasets; the mAP values are 2.02% and 0.71% higher than those of the state-of-the-art model on the OPIXray and SIXray datasets. Compared with the existing one-stage prohibited items detection network, our model can achieve an

optimal detection performance. Especially for the small target category "Straight Knife" in OPIXray, which faces the problem of obscuration and small scale, and its aspect ratio is extremely uneven, our model achieves an 8.88% improvement compared with POD-y. The above experimental results fully demonstrate that our proposed method is effective and efficient.

4.3. Comparing with Different Attention

To verify the effectiveness of the improved attention mechanisms in this paper, we compare the mainstream attention mechanisms, including the SE [51], GAM [52], CA [35] and PSA [53] attention mechanisms. The specific results are shown in Table 3, below.

Table 3. Comparison of different attention mechanism modules.

Model	MAP	GFLOPs	Parameters (M)
YoloX	90.45	155.331	54.152
YoloX + SE	90.75	156.017	54.383
YoloX + GAM	89.87	218.954	88.624
YoloX + CBAM	91.18	156.013	54.383
YoloX + CA	91.49	156.032	54.342
YoloX + DOAM	92.24	175.362	54.290
YoloX + MCA (ours)	92.36	156.037	54.382

It is obvious that our method performs better on the OPIXray dataset compared to the other methods, with results 2.11%, 1.81%, 2.69%, 1.38%, 1.17% and 0.21% higher than the other attention mechanisms, respectively. We also compare DOAM, an attention mechanism for contraband detection, and see that our model is 0.12% more accurate than DOAM, with a smaller number of computations and parameters than DOAM. It can be seen that our model maintains a high level of detection accuracy and speed without a significant increase in the number of computations and parameters.

4.4. Comparing with Different Receptive Field Enhancement Module

We further verify the effect of our multi-scale smoothed atrous convolution (SAC). As we can see in Table 4, we compare different receptive field enhancement modules include ASPP [54] and RFB [55]. Our method shows an improvement of 1.80% and 0.34% over the ASPP and RFB modules.

Table 4. Comparing with different receptive field modules.

Method	Category					mAP	FLOPs	Paras (m)	FPS
	FO	ST	SC	UT	MU				
Baseline	91.84	77.53	97.89	89.22	92.79	89.85	155.331	54.152	89.834
Baseline + ASPP	91.74	78.42	97.87	91.36	92.09	90.03	238.416	100.955	74.567
Baseline + RFB	89.83	86.04	99.3	87.4	94.91	91.49	209.604	83.812	79.058
Baseline + SAC (our)	89.71	88.68	99.32	86.67	94.29	91.83	233.403	96.834	74.350

To better show the superiority of our proposed model, we plot the P-R curves for different receptive field modules, as shown in Figure 9. The P-R curve of our module is closer to the upper right position compared to the other models, which means that our SAC has a better performance.

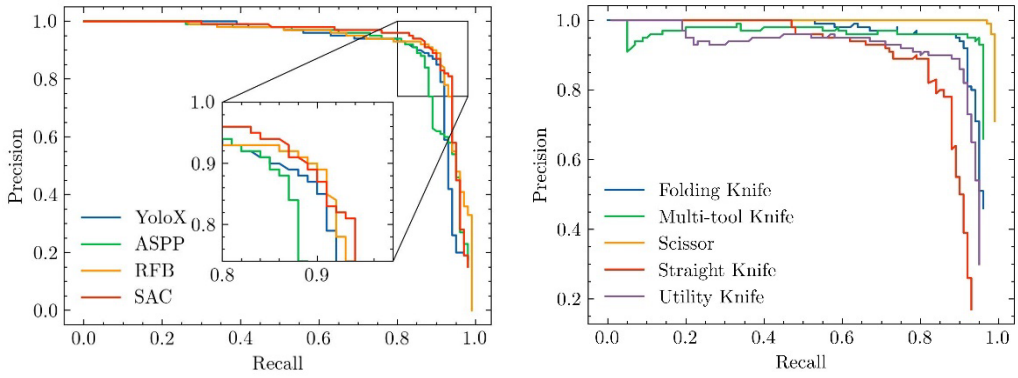


Figure 9. Comparison of P-R curves of different receptive field enhancement modules.

4.5. Ablation Study

To verify the effect of each module on the model performance, we perform ablation experiments on the OPIXray and SIXray datasets. The results are shown in Table 5. We compare the mAP of the model with different combinations of components. The same parameters were used for all the experiments performed in the ablation study to ensure the validity of the comparison. The SD-SIoU increases the mAP of the baseline from 89.85% to 91.76% and 89.80% to 90.94% on OPIXray and SIXray, respectively. This result shows that the SD-SIoU has considerably improved the detecting performance. Then, we split the material-aware path aggregation network into SAC and MCA, which represent the Smoothed Atrous Convolution and Material-aware Coordinate Attention. The MCA increases the mAP of the baseline with SD-SIoU by 0.29% on OPIXray and 0.18% on SIXray. SAC increases the mAP of the baseline with SD-SIoU by 0.60% on the OPIXray and 0.33% on the SIXray. The experiments shows that the SAC and MCA modules are helpful for the model to detect contraband accurately. Finally, when all the methods are used together, our model mAP achieves 92.65%and 91.31%. These are 2.80% and 1.51% higher than the YoloX baseline on OPIXray and SIXray, respectively. Each method can improve performance individually, and combining these methods results in the optimal performance. It is worth mentioning that the improvement on the OPIXray dataset is greater than on the SIXray dataset. The main gap is in the SD-SIoU section. It will be further investigated in the following.

Table 5. Ablation study on OPIXray and SIXray.

SD-SIoU	MCA	SAC	mAP (%)		GFLOPs	Parameters
			OPIXray	SIXray		
			89.85	89.80	156.011	54.209
√			91.76	90.94	156.011	54.209
√	√		92.05	91.12	156.037	54.385
√		√	92.36	91.27	256.835	109.284
√	√	√	92.65	91.31	256.860	109.460

To further visualize the effectiveness of our SD-SIoU, we perform detailed ablation experiments on the SD-SIoU part, which we illustrate by two parts of the mAP and loss function curves.

As can be seen in Table 6, we compared the mAP of the SIoU loss function under different conditions. ON denotes the optimized normalized curve; LSside denotes the long–short side decoupling module. "Decoupling" means the category information embedding module. The optimized normalized curve improves the mAP of the model by 1.38% and 0.95%, which means this normalized method can improve the convergence results of the model. It is worth noting that, when introducing the long–short side decoupling module without the category information embedding module, the accuracy of the model

decreases by 0.16% and 0.21%. The reason for this phenomenon is that there is a serious maldistribution after the construction of the long–short side shape loss. The weight of the long-side loss is not balanced with the weight of the short-side loss. Therefore, we continue to add the category length ratio decoupling module. It increases the mAP by 0.69% and 0.50% and achieves higher AP detection performance.

Table 6. Ablation study for OD-SIoU.

Method	mAP (%)	
	OPIXray	SIXray
SIoU	89.85	89.70
SIoU-ON	91.23	90.65
SIoU-ON-LSside	91.07	90.44
SIoU-NO-LSside-Decoupled	91.76	90.94

We recorded the shape loss curves and long–short loss curves of the SD-SIoU under different conditions. Since the loss data under different conditions varied widely and had small fluctuations, we normalized and denoised all the curves and indicated their validity by observing the decreasing trend of loss. The specific images are shown in the Figure 10.

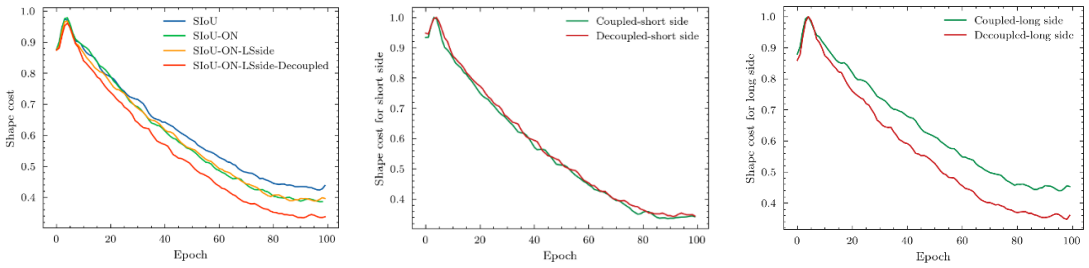


Figure 10. Ablation experiments on loss reduction curves for SD-SIoU.

The first figure shows the SD-SIoU loss curves under the ablation experiment. The loss value drops lower after improving the normalization function of the shape loss factor, but the trend is almost the same at the beginning of the training. This is because our new normalization method still has a good gradient in the late training period, while the gradient of the traditional normalization method is not significant in that period. We also find that the downward trend does not change significantly after adding the long–short side decoupling module, but there is a significant improvement after adding the category information embedding module. To address this issue, we conduct more detailed experiments.

Figure 10 splits the long side and short side from the shape loss. This represents the long-side loss and short-side loss before and after adding the category information embedding module. We can see that the addition of this module directly affects the decreasing trend of the long-side loss, while the decreasing trend of the short edge does not change significantly. This means that adding the classification module can effectively improve the convergence of the long side without affecting the short-side loss. In other words, this module alleviates the problem of uneven weights between the long–short sides.

Finally, we use the model proposed in this paper for visual inspection of the OPIXray and SIXray datasets, as shown in Figure 11 below.

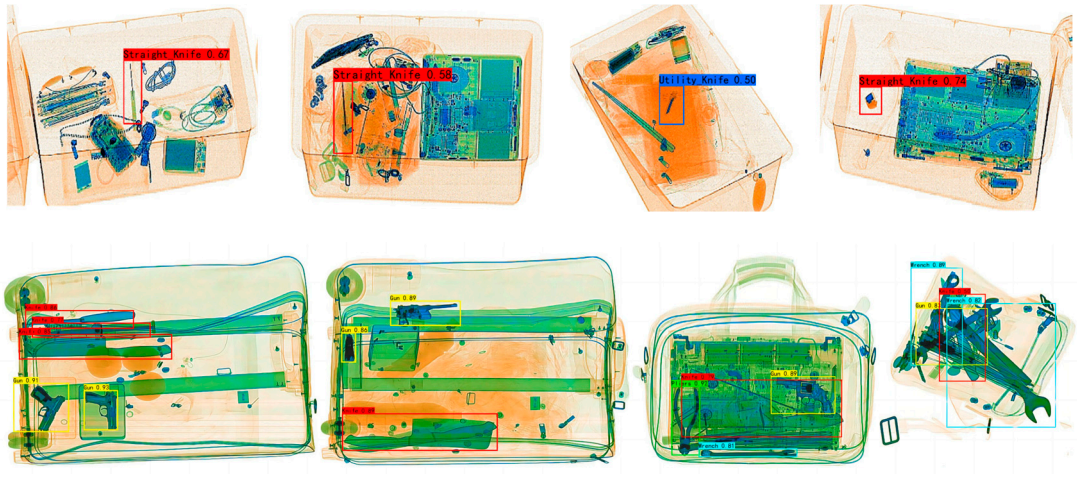


Figure 11. Detection performance in OPIXray and SIXray. The first line is the detection performance of OPIXray and the second line is that of the SIXray.

5. Conclusions

In this paper, a new feature extraction network is designed considering the specific physical characteristics of X-ray images. For the X-ray contraband multi-scale problem, a multi-scale smoothing atrous convolution module is designed to capture multi-scale contraband features by acquiring different sizes of the receptive field. For the occlusion and weak textural information in X-ray contraband images, we design a material-aware coordinate attention mechanism to enhance the material features' extraction ability in obscured X-ray images. In addition, an improved SIoU was designed, named SD-SIoU, which addresses the problem of inconsistent aspect ratios in contraband images. Through a large number of experiments and visualization results, we determine that the feature extraction and enhancement strategies proposed in this paper can effectively strengthen the ability of the model to detect contraband. Its validity is reflected in the evaluation index mAP. Our experimental results, based on the OPIXray and SIXray datasets, show that our method achieves an average accuracy of 92.65% and 91.31%, with a computational volume of 256.86G for 109.46M parameters, respectively. From the quantitative point of view, the proposed method has excellent performance in the field of contraband detection. The comparison results show that the method outperforms other contraband detection methods.

Author Contributions: Methodology, N.X.; software, Z.G.; validation, N.X. and Z.G.; writing—original draft preparation, Z.G.; writing—review and editing, Z.G. and Y.X.; visualization, Y.X.; supervision, N.X. and L.X.; funding acquisition, L.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Chongqing Province of China, grant number CSTB2022NSCQ-MSX0786 and the Natural Science Foundation of Chongqing Province of China, grant number CSTB2022NSCQ-MSX1477.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Michel, S.; Koller, S.M. Computer-based training increases efficiency in X-ray image interpretation by aviation security screeners. In Proceedings of the 2007 41st Annual IEEE International Carnahan Conference on Security Technology, Ottawa, ON, Canada, 8–11 October 2007.
2. Lin, T.-Y.; Maire, M. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
3. Everingham, M.; Van Gool, L. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
4. Thammasorn, P.; Oztan, B. Augmenting data with GANs for firearms detection in cargo X-ray images. In Proceedings of the Anomaly Detection and Imaging with X-rays (ADIX) VII, Orlando, FL, USA, 3 April–13 June 2022.

5. Chang, A.; Zhang, Y. Detecting prohibited objects with physical size constraint from cluttered X-ray baggage images. *Knowl. Based Syst.* **2022**, *237*, 107916. [CrossRef]
6. Velayudhan, D.; Hassan, T. Baggage threat recognition using deep low-rank broad learning detector. In Proceedings of the 2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON), Palermo, Italy, 14–16 June 2022.
7. Velayudhan, D.; Hassan, T. Recent advances in baggage threat detection: A comprehensive and systematic survey. *ACM Comput. Surv.* **2022**, *55*, 1–38. [CrossRef]
8. Ge, Z.; Liu, S. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
9. Wei, Y.; Tao, R. Occluded prohibited items detection: An X-ray security inspection benchmark and de-occlusion attention module. In Proceedings of the Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020.
10. Miao, C.; Xie, L. 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, Long Beach, CA, USA, 15–20 June 2019.
11. Turcsany, D.; Mouton, A. Improving feature-based object recognition for X-ray baggage security screening using primed visualwords. In Proceedings of the 2013 IEEE International conference on industrial technology (ICIT), Cape Town, South Africa, 25–28 February 2013.
12. Zhang, N.; Zhu, J. A study of X-ray machine image local semantic features extraction model based on bag-of-words for airport security. *Int. J. Smart Sens. Intell. Syst.* **2015**, *8*, 45–64. [CrossRef]
13. Akçay, S.; Kundegorski, M.E. Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016.
14. Li, J.; Liu, Y. Segmentation and Attention Network for Complicated X-Ray Images. In Proceedings of the 2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC), Zhanjiang, China, 16–18 October 2020.
15. He, K.; Gkioxari, G. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
16. Zhang, Y.; Kong, W. On using XMC R-CNN model for contraband detection within X-ray baggage security images. *Math. Probl. Eng.* **2020**, *2020*, 1823034. [CrossRef]
17. Wang, B.; Zhang, L. Towards real-world prohibited item detection: A large-scale X-ray benchmark. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
18. Tao, R.; Wei, Y. 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, Montreal, QC, Canada, 10–17 October 2021.
19. Ma, C.; Zhuo, L. EAOD-Net: Effective anomaly object detection networks for X-ray images. *IET Image Process.* **2022**, *16*, 2638–2651. [CrossRef]
20. Nguyen, H.D.; Cai, R. Towards More Efficient Security Inspection via Deep Learning: A Task-Driven X-ray Image Cropping Scheme. *Micromachines* **2022**, *13*, 565. [CrossRef]
21. Ma, C.; Zhuo, L. Occluded prohibited object detection in X-ray images with global Context-aware Multi-Scale feature Aggregation. *Neurocomputing* **2023**, *519*, 1–16. [CrossRef]
22. Gaus, Y.F.A.; Bhowmik, N. Evaluating the transferability and adversarial discrimination of convolutional neural networks for threat object detection and classification within X-ray security imagery. In Proceedings of the 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019.
23. Hassan, T.; Shafay, M. Meta-transfer learning driven tensor-shot detector for the autonomous localization and recognition of concealed baggage threats. *Sensors* **2020**, *20*, 6450. [CrossRef]
24. Li, D.; Hu, X. A GAN based method for multiple prohibited items synthesis of X-ray security image. *Optoelectron. Lett.* **2021**, *17*, 112–117. [CrossRef]
25. Hassan, T.; Akcay, S. Tensor pooling-driven instance segmentation framework for baggage threat recognition. *Neural Comput. Appl.* **2022**, *34*, 1239–1250. [CrossRef]
26. Liu, D.; Tian, Y. Handling occlusion in prohibited item detection from X-ray images. *Neural Comput. Appl.* **2022**, *34*, 20285–20298. [CrossRef]
27. Yan, Y.; Li, J. Anchor-free person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
28. Bochkovskiy, A.; Wang, C.-Y. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
29. Liu, S.; Qi, L. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
30. Cao, F.; Bao, Q. A survey on image semantic segmentation methods with convolutional neural network. In Proceedings of the 2020 International Conference on Communications, Information System and Computer Engineering (CISCE), Kuala Lumpur, Malaysia, 3–5 July 2020.
31. Wang, Z.; Ji, S. Smoothed dilated convolutions for improved dense prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018.

32. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
33. Su, C.; Wu, X. Restoration of turbulence-degraded images using the modified convolutional neural network. *Appl. Intell.* **2022**, *53*, 5834–5844. [CrossRef]
34. He, K.; Zhang, X. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
35. Hou, Q.; Zhou, D. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
36. Lee, H.; Kim, H.-E. Srm: A style-based recalibration module for convolutional neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
37. Gevorgyan, Z. SIoU Loss: More Powerful Learning for Bounding Box Regression. *arXiv* **2022**, arXiv:2205.12740.
38. Zheng, Z.; Wang, P. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
39. Rezatofighi, H.; Tsoi, N. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
40. Xue, Q.; Lin, H. FCDM: An Improved Forest Fire Classification and Detection Model Based on YOLOv5. *Forests* **2022**, *13*, 2129. [CrossRef]
41. Liu, B.; Luo, H. An Improved Yolov5 for Multi-Rotor UAV Detection. *Electronics* **2022**, *11*, 2330. [CrossRef]
42. Guo, Y.; Chen, S. LMSD-YOLO: A Lightweight YOLO Algorithm for Multi-Scale SAR Ship Detection. *Remote Sens.* **2022**, *14*, 4801. [CrossRef]
43. Yang, X.; Zhao, J. Detection of River Floating Garbage Based on Improved YOLOv5. *Mathematics* **2022**, *10*, 4366. [CrossRef]
44. Liu, Z.; Lin, Y. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
45. Lin, T.-Y.; Goyal, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
46. Qiao, S.; Chen, L.-C. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
47. Shao, F.; Liu, J. Exploiting foreground and background separation for prohibited item detection in overlapping X-ray images. *Pattern Recognit.* **2022**, *122*, 108261. [CrossRef]
48. Wei, Y.; Wang, Y. CFPA-Net: Cross-layer Feature Fusion And Parallel Attention Network For Detection And Classification of Prohibited Items in X-ray Baggage Images. In Proceedings of the 2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS), Xi'an, China, 7–8 November 2021.
49. Wang, M.; Du, H. Material-aware Cross-channel Interaction Attention (MCIA) for occluded prohibited item detection. *Vis. Comput.* **2022**. [CrossRef]
50. Zhao, C.; Zhu, L. Detecting Overlapped Objects in X-ray Security Imagery by a Label-Aware Mechanism. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 998–1009. [CrossRef]
51. Hu, J.; Shen, L. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
52. Liu, Y.; Shao, Z. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. *arXiv* **2021**, arXiv:2112.05561.
53. Zhang, H.; Zu, K. EPSANet: An efficient pyramid squeeze attention block on convolutional neural network. In Proceedings of the Asian Conference on Computer Vision, Macau, China, 4–8 December 2022.
54. Pu, R.; Ren, G. Autonomous Concrete Crack Semantic Segmentation Using Deep Fully Convolutional Encoder–Decoder Network in Concrete Structures Inspection. *Buildings* **2022**, *12*, 2019. [CrossRef]
55. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.