electronics

MDPI

Article

X-ray Detection of Prohibited Item Method Based on Dual Attention Mechanism

Ying Li [1], Changshe Zhang [2], Shiyu Sun [3] and Guangsong Yang [2,*]

[1] Chengyi College, Jimei University, Xiamen 361021, China
[2] School of Ocean Information Engineering College, Jimei University, Xiamen 361021, China
[3] School of Computer Engineering College, Jimei University, Xiamen 361021, China
* Correspondence: gsyang@jmu.edu.cn

Abstract: Prohibited item detection plays a significant role in ensuring public safety, as the timely and accurate identification of prohibited items ensures the safety of lives and property. X-ray transmission imaging technology is commonly employed for prohibited item detection in public spaces, producing X-ray images of luggage to visualize their internal contents. However, challenges such as multiple object overlapping, varying angles, loss of details, and small targets in X-ray transmission imaging pose significant obstacles to prohibited item detection. Therefore, a dual attention mechanism network (DAMN) for X-ray prohibited item detection is proposed. The DAMN consists of three modules, i.e., spatial attention, channel attention, and dependency relationship optimization. A long-range dependency model is achieved by employing a dual attention mechanism with spatial and channel attention, effectively extracting feature information. Meanwhile, the dependency relationship module is integrated to address the shortcomings of traditional convolutional networks in terms of short-range correlations. We conducted experiments comparing the DAMN with several existing algorithms on datasets containing 12 categories of prohibited items, including firearms and knives. The results show that the DAMN has a good performance, particularly in scenarios involving small object detection, detail loss, and target overlap under complex conditions. Specifically, the detection average precision of the DAMN reaches 63.8%, with a segmentation average precision of 54.7%.

Keywords: deep learning; object detection; prohibited item detection; spatial attention; channel attention

Citation: Li, Y.; Zhang, C.; Sun, S.; Yang, G. X-ray Detection of Prohibited Item Method Based on Dual Attention Mechanism. *Electronics* **2023**, *12*, 3934. https://doi.org/10.3390/electronics12183934

Academic Editors: Fabio Mendonca, Morgado Dias and Sheikh Shanawaz Mostafa

Received: 27 August 2023
Revised: 12 September 2023
Accepted: 16 September 2023
Published: 18 September 2023

1. Introduction

Prohibited item detection is a crucial means of ensuring public safety. With the rapid development of public transportation, such as civil aviation aircraft, high-speed trains, subways, and metro systems, convenient and efficient travel options are provided to people. However, this progress also brings security risks. The significant flow of passengers exposes public transport areas to serious safety hazards. Any accidents occurring in such areas can pose severe threats to both lives and property. Therefore, conducting prohibited item detection in public places is of the utmost importance.

Currently, prohibited item detection primarily relies on X-ray security screening technology. X-rays can penetrate objects and generate perspective images, enabling a clear observation and examination of items within the baggage. This effectively prevents situations where passengers may carry restricted items such as knives, firearms, ammunition, and lighters that could pose threats to passengers or transportation systems. However, the current practice of manually discerning prohibited items from X-ray images still heavily relies on the professional knowledge, experience, and condition of security personnel [1]. Prolonged work hours can lead to visual fatigue among security personnel, thereby increasing the risk of false positives and false negatives. Furthermore, the workload and intensity of security screening are substantial, especially during periods of high passenger flow. As the volume of items to be inspected increases, security personnel find it challenging to make

quick judgments. This can result in passenger delays and congestion and may even lead to safety and public health concerns. To enhance the efficiency and accuracy of prohibited item detection, it is imperative to further develop automated detection technology. By integrating advanced deep learning techniques, the automatic detection and recognition of prohibited items within X-ray images can be achieved. This will alleviate the burden on security personnel, enhance the speed and accuracy of security checks, and thus better ensure the safety of public spaces.

Over the past few years, deep learning techniques have found widespread applications across various domains [2]. In the field of prohibited item detection, researchers have attempted to construct intelligent convolutional neural networks for prohibited item recognition. The aim is to reduce both false negatives and false positives while simultaneously enhancing the speed and accuracy of prohibited item detection. However, traditional deep-learning-based image object detection methods primarily object ordinary optical images. As the imaging principles of X-ray penetrating objects differ from those of ordinary optical images, and the scenarios for danger detection are distinct, existing methods cannot be directly applied to prohibited item detection in X-ray images. Firstly, when multiple objects overlap, X-rays penetrate these objects, resulting in overlapped image contours and a more complex depiction of the image contours. This complexity makes differentiating between objects and accurately extracting boundaries more challenging. Secondly, the placement angles and positions of baggage items are different, leading to varying imaging angles of objects, which drastically change from different perspectives, which increase the detection difficulty and need algorithms to handle objects from various angles and poses. Furthermore, while X-ray imaging easily captures object contour information, it may cause a loss of surface detail features. This loss of detail might lead to difficulties in distinguishing specific features of objects, thereby affecting the accurate detection of prohibited items. Moreover, X-ray imaging cannot directly capture object color information. Instead, it uses measurements of high- and low-energy X-rays to determine an object's atomic number, distinguishing between organic and inorganic substances and, finally, coloring the object. So, other features instead of color need to be used to detect prohibited items. Additionally, for the purpose of concealment, most prohibited items are typically small and hidden in intricate environments, posing a more challenging task for detection. To recap, X-ray transmission imaging faces challenges in prohibited item detection, including multiple object overlaps, varying angles, loss of detail features, and small objects. Addressing these challenges requires the targeted design and optimization of algorithms to improve the accuracy and efficiency of prohibited item detection.

This paper presents a dual attention mechanism network (DAMN) designed for X-ray prohibited item detection by incorporating a fused dual attention mechanism. The central goal is to extract crucial information from key target areas, thereby elevating the precision and efficiency of X-ray prohibited item detection. The primary components of this paper are outlined as follows:

Section 1 provides a concise overview of X-ray prohibited item detection. In Section 2, the focus is on the exploration of X-ray prohibited item detection methods based on deep learning techniques. Section 3 discusses the principle and structure of the proposed dual attention mechanism for X-ray prohibited item detection (a DAMN). Section 4 conducts experimental comparisons and performance evaluations of the proposed DAMN algorithm with classical algorithms such as YOLOv5, RetinaNet, Mask R-CNN, Cascade Mask R-CNN, and SDANet for X-ray prohibited item detection. Section 5 concludes with a summary of the findings and an outlook for future research directions.

2. Related Work

Traditional machine learning methods for X-ray prohibited item image detection rely on manual feature extraction and classifier-based classification. The localization information is obtained using a sliding window approach [3]. While effective for processing small batches of data with prominent features, these methods exhibit limited accuracy and

generalization capabilities when it comes to recognizing complex and dynamic prohibited items. In contrast, the evolution of deep learning techniques has led to the application of methods like convolutional neural networks (CNNs) in X-ray image detection. These deep learning approaches excel by enabling nonlinear feature learning, rendering them suitable for large-scale datasets and intricate background scenarios. The inherent strengths of deep learning methods lie in their potent representational capacity, adaptability, and flexibility, which manifest remarkably in tasks such as classification and detection, particularly within intricate environments. As a result, deep learning techniques have been extensively applied in the field of X-ray prohibited item detection, markedly enhancing the accuracy and robustness of detection. This development provides an effective means to bolster security measures in public spaces.

In 2016, Akcay et al. [4] pioneered the integration of deep learning into X-ray baggage security image detection. They addressed the image classification problem using transfer learning and the AlexNet architecture, demonstrating remarkable detection performance and robustness. Subsequently, Mery et al. [5] explored various techniques, including classical methods, the bag-of-words model, sparse representation, and deep learning. They achieved high detection performance on both the AlexNet and GoogleNet architectures. In 2017, Akcay et al. [6] further employed transfer learning and compared the detection performance of CNN models based on sliding windows with detection models based on candidate regions on the ImageNet dataset. They found that the candidate-region-based model exhibited a higher detection accuracy. In 2018, Akcay et al. [7] conducted prohibited item detection experiments using various deep learning classification models and the bag-of-words model. The results demonstrated that, compared to traditional machine learning approaches, deep learning methods were better suited for X-ray prohibited item detection, showcasing greater practicality and superiority. In 2019, Jinyi Liu et al. [8] introduced a color-based foreground–background segmentation method that addressed brightness differences between the foreground and background, reducing color information loss and enhancing the classification and localization performance of prohibited item detection. Also, in the same year, Neelanjan Bhowmik et al. [9] employed the TIP (texture inpainting with prior) method to synthesize images and utilized the R-CNN and ResNet-101CNN architectures to enhance the detection capabilities of overlapping prohibited items. In 2020, Subramani [10] proposed the RetinaNet network, which tackled the challenges of missed detections and false alarms for small object prohibited items. Similarly, in the same year, Zhigang Su et al. [11] utilized the ASPC (atrous spatial pyramid convolution) module for multiscale feature extraction within the shallow layers of the VGG16 architecture, effectively addressing the issue of detail information loss. In 2021, Jian Gu [12] improved the Canny edge detection algorithm to remove background noise interference and incorporated the SPP (spatial pyramid pooling) module into the YOLOv3 network, enhancing the detection capabilities for small object prohibited items. In 2022, Yishan Dong et al. [13] introduced a weighted bounding box fusion algorithm into the YOLOv5 network, enhancing the detection accuracy of X-ray prohibited item datasets. In 2023, Song Li et al. [14] employed the MixUP data augmentation method to improve the clarity of overlapping image contours. They also utilized the MobileNet ViTv3 block to extract more features on a global scale, thereby reducing instances of missed detection for overlapping prohibited items. Overall, deep learning methods have demonstrated significant potential and advantages in X-ray prohibited item detection, continuously driving advancements in the field, as summarized in Table 1.

Table 1. Summary of X-ray prohibited item detection studies based on deep learning techniques.

Dataset	Network Framework	Method	Contribute	References	Year
Firearms (17,419)	AlexNet	CNN using transfer learning	Higher detection performance and robustness	Akcay, etc. [4]	2016
GDXray (1950)	AlexNet GoogleNet	Compared to traditional machine learning methods	Deep learning methods achieve higher detection accuracy	Mery, etc. [5]	2016
Baggage security (11,627)	SW-CNN R-CNN Faster-RCNN	CNN for transfer learning	Improve real-time applicability of detection	Akcay, etc. [6]	2017
Baggage security (3080)	SSD	Design SSD detectors and tracking managers	Improve the multiobject detection accuracy	Han [15]	2018
Baggage security (12,700)	SSD	New feature maps generated by MF are fed into ATM and DCM	Better detection accuracy for multiple obstructed prohibited items	Zhang, etc. [16]	2020
Baggage security (550)	SSD	Lightweight network integration into feature pyramid layer CBAM	Better small object detection accuracy	Zhang, etc. [17]	2020
Baggage security (2026)	YOLOv5	Incorporating adaptive feature fusion (ASFF) and attention mechanism (CBAM)	Higher accuracy of small object detection	Ren [18]	2021
SIXray (8900)	YOLOv4	Adding attention mechanism to hollow dense convolutional modules	Better detection accuracy for multiple obstructed prohibited items	Mu, etc. [19]	2021
SIXray_OD (8718)	Faster-RCNN	Introducing dilation convolution to construct a multiscale detection network	Improving the detection accuracy of small object prohibited items	Kang, etc. [20]	2022
OPIXray (8885) HiXray (45,364), etc.	YOLOv5	Add the CBAM attention mechanism	Improve detection capabilities for overlapping prohibited items	Dong, etc. [13]	2022
HiXary (9565) CLCXray (45,364)	YOLOv7	Add Mobile NetViTv3 block	Reduce the missed detection rate of overlapping prohibited items	Li, etc. [14]	2023

3. Proposed Method

3.1. Network Architecture

We proposed a deep-learning-assisted X-ray prohibited item detection method (a DAMN) aimed at achieving enhanced detection performance and accuracy in this domain. The fundamental architecture comprises a backbone network, a feature pyramid network (FPN) [21], attention mechanisms, and a region proposal network (RPN), as illustrated in Figure 1.

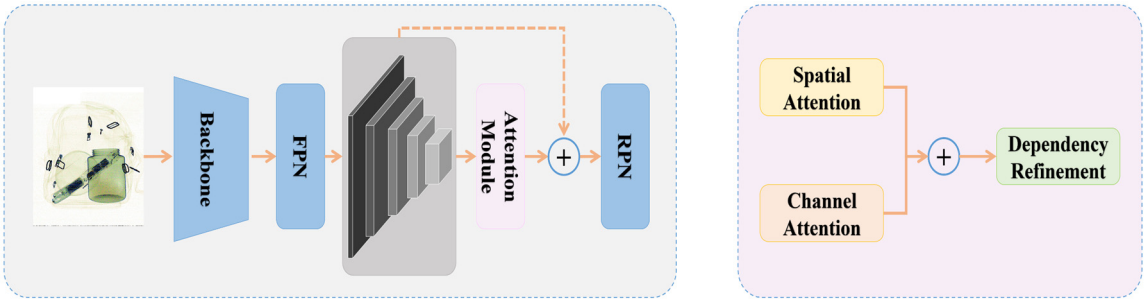


Figure 1. X-ray detection of prohibited item network based on dual attention mechanism.

The backbone network is ResNet-101-FPN, a type of deep residual network comprising 101 layers, utilized for extracting rich high-level features from input X-ray images, which was constructed by stacking residual blocks, employing skip connections and residual connections to mitigate the gradient vanishing problem during deep network training,

effectively capturing image details and semantic information. Furthermore, to enhance the network's ability to represent multiscale features, a feature pyramid network (FPN) was introduced. FPN establishes connections between feature maps at different levels, facilitating the fusion of multiscale information. This multiscale feature fusion strategy is instrumental in addressing the diversity of prohibited items' images, encompassing variations in category and size.

We employed a dual attention module combined with spatial attention and channel attention modules in parallel. The spatial attention module aims to learn spatial dependencies within the feature map, enhancing the precise localization of prohibited items by emphasizing crucial regions. On the other hand, the channel attention module focuses on the interchannel correlations within the feature map, bolstering the network's discriminative capability toward prohibited item detection.

To further enhance the network's performance in prohibited item detection tasks, a dependency relationship optimization module (DROM) was introduced, which aims to extract relationships between channels within the feature map and integrate these relationships into the positional representation of feature information. This strategy effectively captures the detailed features and contextual information of prohibited items, resulting in a significant improvement in detection accuracy.

In the DAMN, we adopted cross-entropy loss [22] as the optimization objective during the training process. The core capability of this loss function lies in its ability to precisely assess the disparity between the model's predicted outcomes and the actual labels, thereby propelling the model to learn and optimize toward greater accuracy. By incorporating the cross-entropy loss into the model's training, the convergence speed is increased, and the classification performance is also improved.

3.2. Spatial Attention Module

In the computer vision field, there are two types of attention mechanisms, i.e., channel attention and spatial attention. The spatial attention mechanism employs a two-dimensional weighting strategy to compress the feature map along the channel dimension and assign independent weights to each pixel, as shown in Figure 2.

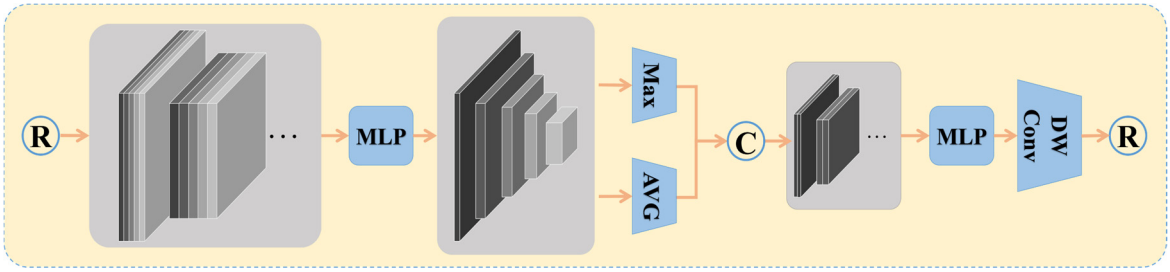


Figure 2. Spatial attention module.

In the spatial attention module, the first step involves standardizing and fusing the multiscale features obtained from the feature pyramid. The feature maps with five different sizes, [4, 256, 128, 120], [4, 256, 64, 60], [4, 256, 32, 30], [4, 256, 16, 15], and [4, 256, 8, 8], are adjusted to a unified size, resulting in a feature tensor $f \in R^{b \times c \times l \times w \times h}$. Here, b represents the batch size, c is the number of channels, w and h are the width and height, and $l = 5$ indicates the presence of five scales of feature maps. Next, information is extracted from feature maps of different dimensions using the nonlinear mappings of the multilayer perceptron (MLP), which can reveal high-level semantics hidden within the feature maps. Then, two distinct feature maps are generated through average pooling and max pooling to describe spatial positional information in the image, aiding the model in better understanding the relationships between different regions. Finally, the spatial disparity information is extracted. These two feature maps are concatenated and operated upon using depthwise separable convolution (DWConv) in combination with the MLP. This

operation further extracts spatial disparity information. The design of DWConv combined with the MLP captures finer-grained spatial variations.

The specific process is illustrated by Equations (1)–(3), where $\hat{F}_{ci}$ represents the features processed by the spatial attention mechanism, and F_i denotes the multiscale features output from the feature pyramid RPN, which includes feature maps of different scales denoted by $F_i \in R^{b\times c\times w\times h}, i \in [0,4]$.

$$\hat{F}_{ci} = \text{concat}(\text{resize}(F_i)), \forall i \tag{1}$$

$$\hat{F}_{ci} = \text{concat}\left(\text{Max}\left(\text{MLP}\left(\hat{F}_{ci}\right)\right), \text{Avg}\left(\text{MLP}\left(\hat{F}_{ci}\right)\right)\right) \tag{2}$$

$$\hat{F}_{ci} = \text{resize}\left(\text{DWConv}\left(\text{MLP}\left(\hat{F}_{ci}\right)\right)\right) \tag{3}$$

3.3. Channel Attention Module

Channel attention is a critical attention mechanism in the field of deep learning, employed to enhance the relationships between different channels within an image (e.g., color or feature channels) to improve a model's performance in specific tasks. In reference [23], channel attention is applied to RGB-D image processing to amplify information in specific channels, thus enhancing the feature representation of images and significantly boosting detection performance. SENet [24] introduced the squeeze-and-excitation module, which initially compresses spatial information and then learns weights to emphasize the importance of different channels. Additionally, CBAM [25] combines channel attention with spatial attention, dynamically weighting feature maps in other media and further enhancing these feature maps at specific positions and scales, ultimately improving the performance and generalization of convolutional neural network (CNN) models.

The channel attention module is a one-dimensional weighting method, as shown in Figure 3, which operates by compressing the spatial dimension of the feature map while allocating distinct weights to each channel. The core concept is to accentuate the feature information of crucial channels, enhancing the representation of essential characteristics.

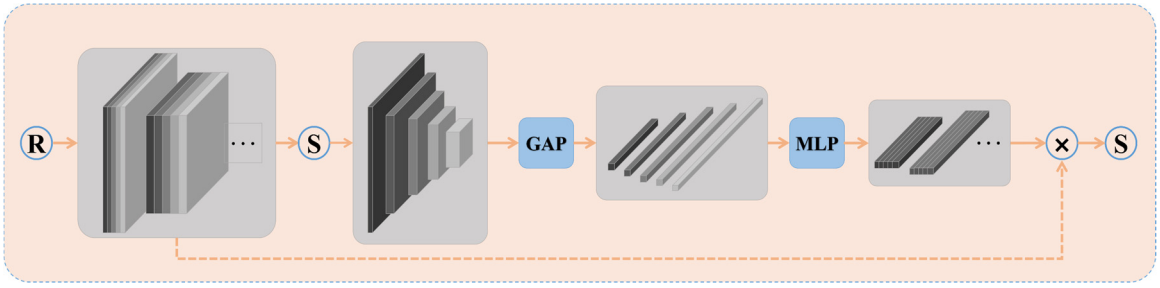


Figure 3. Channel attention module.

After obtaining multiscale information from the feature pyramid network (FPN), the channel attention module first introduces scale standardization and feature fusion (similar to 3.2), aligning and merging feature maps of different scales. Subsequently, global average pooling (GAP) is applied to the fused feature maps for global information extraction. GAP averages the features of each channel, reducing the spatial dimensions of the feature map to one dimension, which diminishes reliance on spatial information and reduces redundant data. The channel attention mechanism focuses the network more on the channel-level significant features, thereby enhancing the network's performance in object detection tasks. Next, through multiple fully connected layers, weight adjustments, and activation functions, the multilayer perceptron (MLP) learns complex mapping relationships of input features, further extracting abstract and advanced prohibited-item-related characteristics hidden within channel interactions.

Finally, the feature $\hat{F}_{si}$ is obtained through the processing of the channel attention module. This process can be represented using Equations (4)–(6). Here, F_i represents the multiscale features' output from the feature pyramid RPN.

$$\hat{F}_{si} = \text{concat}(\text{resize}(F_i)), \forall i \tag{4}$$

$$\hat{F}_{si} = F_i \times \text{MLP}(\text{GAP}(\text{sum}(\hat{F}_{si}))) \tag{5}$$

$$\hat{F}_{si} = \text{sum}(\hat{F}_{si}) \tag{6}$$

3.4. Dependency Refinement

Based on the foundation of channel and spatial attention, the feature map was further refined using the dependency relationship optimization module, as shown in Figure 4. This process generates highly discriminative feature maps capable of capturing long-range dependencies within the image and aids in achieving a more accurate understanding of the semantic content within prohibited items' images.

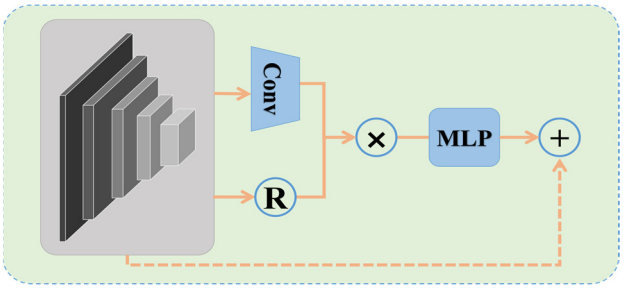


Figure 4. Dependency relationship module.

Firstly, by integration of channel and spatial attention mechanisms, we obtained the aggregated feature F_{sci}, which not only emphasizes crucial channels and important spatial positions but also enhances the feature's expressive capability to a certain extent. Furthermore, the dependency relationship optimization module was introduced with the primary objective of further enhancing the discriminative capability of the feature map. Within this module, a nonlocal representation method [26,27], as described in Equation (8), was employed. By applying the convolutional operations to the resized feature map, it effectively captures long-range dependencies within the image, thus better-expressing feature information. By aggregating global contextual features, the overall image semantics are well known. Moreover, establishing relationships between different channels further reinforces channel correlations, enhancing feature distinctiveness. Finally, we used a fusion module to fuse global contextual features into features across all positions, as shown in Equation (9). Adding MLP-processed features to the original features to fuse global contextual information can help obtain the feature map from global contextual information at all positions, further enhancing its representational capacity.

The entire process can be further described using Equations (7)–(9), where $\hat{F}_i$ represents the features processed by the dependency relationship optimization module, and F_{sci} is the aggregated feature obtained by combining channel and spatial attention.

$$F_{sci} = F_{si} + F_{ci} \tag{7}$$

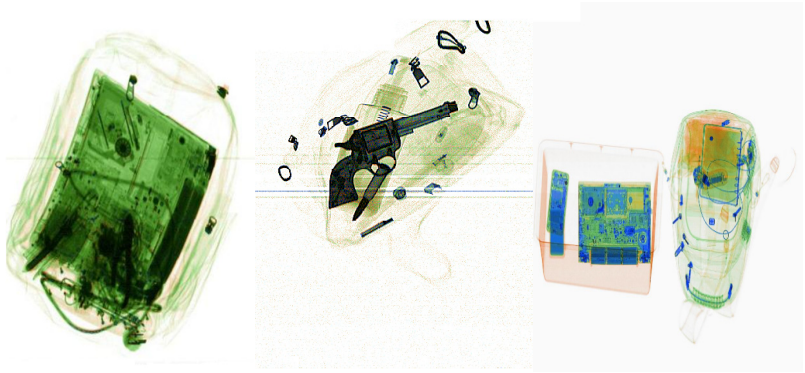
$$\hat{F}_i = \text{resize}(F_{sci}) \times \text{Conv}(F_{sci}) \tag{8}$$

$$\hat{F}_i = F_{sci} + \text{MLP}(\hat{F}_i) \tag{9}$$

4. Experiment and Performance Evaluation

4.1. Dataset

We used the PIDray hazardous material detection dataset [28] as our foundational dataset. The image's resolution was 1020 × 1040. The dataset was formatted in a COCO format and was classified into 29,457 for the training set, 9482 for the easy test set, 3733 for the challenging test set, and 5005 for the hidden test set. The images within the easy test set exclusively featured a single prohibited item, as illustrated in Figure 5a, while the challenging test set comprised images containing multiple prohibited items, as depicted in Figure 5b, and the hidden test set encompassed images intentionally concealing prohibited items, as demonstrated in Figure 5c.



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

Figure 5. Test dataset classification. (**a**) Easy legend: pliers. (**b**) Hard legend: pistol and bullet. (**c**) Hidden legend: winding power bank.

4.2. Experimental Parameter Settings

To ascertain the effectiveness of the DAMN, we chose a few State-of-the-Art object detection algorithms as baseline benchmarks for our experiments, including single-stage YOLOv5, RetinaNet, two-stage Mask R-CNN, Cascade Mask R-CNN, and SDANet.

In YOLOv5, we adopted the YOLOv5s architecture with an initial learning rate of 0.01. The input image size was 640 × 640, and each training batch contained 32 images. We implemented the early stopping strategy and utilized the one-cycle policy for the learning rate scheduling, resulting in training stopping at the 50th epoch. For RetinaNet, we employed a ResNet50 backbone with an initial learning rate of 1×10^{-5}. The input image size was 608 × 608, and each training batch included two images. We applied the reduced learning rate on the plateau strategy to ensure model convergence within 50 epochs and halted training after achieving convergence. Regarding the two-stage networks, we used a ResNet101 backbone pretrained on ImageNet, with a learning rate of 0.02. The input image size was 500 × 500, and each training batch contained four images. In the case of Mask R-CNN, Cascade Mask R-CNN, and SDANet, a step-wise learning rate strategy with warm-up was employed, reducing the learning rate to one-third of its original value at the 25th and 40th epochs.

In the experiments, the "early stop" strategy was applied to all models, which entails halting training when the models exhibit convergence. The training was conducted on two NVIDIA GeForce 1080Ti GPUs, utilizing CUDA version 10.0 and CUDNN version 7.6.5 to ensure consistency and stability within the experimental environment.

4.3. Experimental Results and Analysis

Table 2 presents a comparative analysis of the experimental results for various methods on the PIDray dataset. As shown in Table 2, the DAMN demonstrates superior performance in prohibited item detection. In our experiments, we conducted tests on the easy, hard, and hidden subsets of the PIDray dataset, comparing the detection average precision (detection AP) of six different methods: Mask R-CNN, Cascade Mask R-CNN, SDANet, YOLOv5,

RetinaNet, and the DAMN. Additionally, we computed the segmentation average precision (segmentation AP). In this context, the COCO mmAP (%) [29] was employed to evaluate the method performance. Specifically, the AP signified the average precision calculated across all 12 categories, and the resulting average values were obtained by considering these categories' mean average precision across 10 different IOU thresholds (ranging from 0.5 to 0.95, with an interval of 0.05).

Table 2. Comparison with other methods on the PIDray [28].

Method	Backbone	Detection AP				Segmentation AP			
		Easy	Hard	Hidden	Overall	Easy	Hard	Hidden	Overall
Mask R-CNN	ResNet-101-FPN	64.6	59.0	43.7	55.7	57.5	50.1	35.1	47.6
Cascade Mask R-CNN	ResNet-101-FPN	70.8	63.0	47.0	61.0	59.1	51.4	36.1	48.8
SDANet	ResNet-101-FPN	71.1	64.1	49.5	61.5	59.8	51.0	37.4	49.7
YOLOv5	CSP Darknet53	71.8	67.3	43.3	60.8	-	-	-	-
RetinaNet	ResNet50	64.0	58.6	41.4	54.7	-	-	-	-
DAMN	ResNet-101-FPN	72.7	66.6	52.1	63.8	64.0	58.6	41.4	54.7

The results from Table 2 clearly demonstrate that the DAMN achieved the best AP performance in the easy, hidden, and overall subsets. Of particular note is the DAMN's remarkable performance in the most challenging hidden subset, where its detection AP is 2.6% higher than the second-ranked SDANet. This indicates that the DAMN effectively leverages attention mechanisms to more efficiently extract feature information from concealed items, leading to more accurate detection. Additionally, the DAMN's segmentation performance significantly outperforms the other networks, with an AP on the hidden subset that surpasses the second-ranked SDANet by 4.0%. However, we also observe that in the hard subset, YOLOv5 exhibits the best detection AP, outperforming the DAMN by a marginal 0.7%. This is due to YOLOv5's advantage in detecting objects of various sizes with its multiscale features. Nonetheless, in the deliberately concealed prohibited items of the hidden subset, the DAMN's detection AP improves by an impressive 8.8% compared to YOLOv5. These experimental results definitively establish that the DAMN method achieved a significant breakthrough in prohibited item detection, particularly showcasing exceptional performance in more challenging hidden scenarios.

Based on the Cascade Mask R-CNN [30], we performed four ablation experiments, as shown in Table 3, to analyze the advantages of our modifications compared to other attention modules.

Table 3. Ablation study of the proposed modules.

Method	Channel Attention	Spatial Attention	Dependency Refinement	Detection AP
Cascade Mask R-CNN				0.470
A	√			0.492
B		√		0.497
C	√	√		0.513
D	√	√		0.502
E(DAMN)	√	√	√	0.521

Method A only optimizes the channel attention module by adding a channel attention module of Figure 2 between the FPN and RPN of the Cascade Mask R-CNN to increase the model's attention to different channels or feature maps, thereby helping the model select and highlight the most critical feature channels.

Method B only optimizes the spatial attention module by adding a spatial attention module of Figure 2 between the FPN and RPN of the Cascade Mask R-CNN to enhance

the model's perception of spatial information, which helps the model better recognize and distinguish features at different positions in the input image, thereby improving the detection and classification performance.

Method C combines channel attention and spatial attention for optimization while focusing on different positions and channel features to capture input data information more comprehensively and improve detection performance.

Method D introduces the CBAM [25] structure, in which the spatial attention module and channel attention module are combined serially.

Method E is our proposed DAMN method. Unlike the CBAM, the channel attention module and spatial attention module are parallel in the DAMN. Moreover, the DAMN introduces the dependency module of Figure 4 and further optimizes the dependency relationship between the internal features of the model to improve its performance and understanding of the semantic content.

The results of the ablation experiment are shown in Table 3. As can be seen from the figure, methods A, B, C, D, and E increased detection AP by 2.2%, 2.7%, 4.3%, 3.2%, and 5.1%, respectively. It is worth noting that the DAMN achieved the most significant performance improvement, reaching the highest detection AP of 52.1%. This series of experimental results indicates that integrating spatial attention, channel attention, and dependency optimization with dependency refinement can further improve the detection accuracy of X-ray contraband based on the advantages of the original improvement.

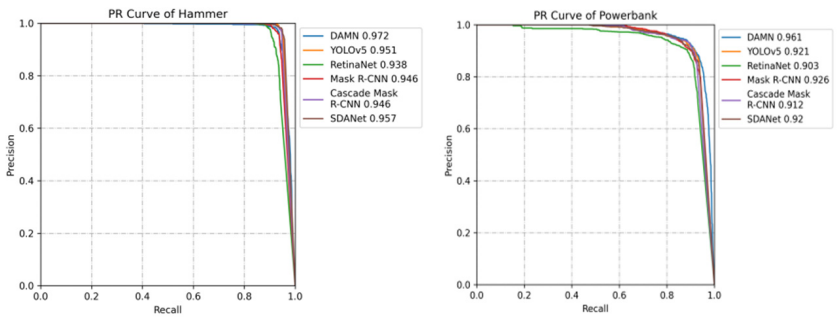
We set the intersection over union (IOU) threshold between the predicted bounding boxes and ground truth boxes to 0.5, meaning that when the IOU is greater than 0.5, the prediction is correct.

We performed several experiments to compare the DAMN model with five existing object detection algorithms on three different datasets, i.e., easy, hard, and hidden.

The specific experimental results are shown in Table 4 and Figures 6–8 below. On the easy dataset, the DAMN exhibits outstanding performance with a mAP of 86.7%. Notably, it excels in detecting common contraband items like the hammer, power bank, and bullet, with mAP scores consistently exceeding 96%.

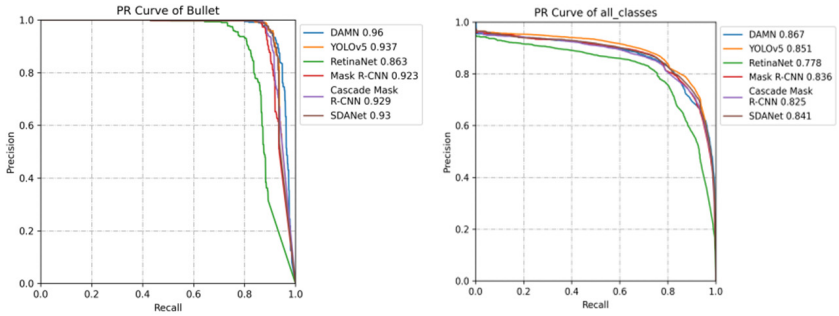
Table 4. Performance comparison of different models.

Method	Backbone	Detection mAP@0.5			
		Easy	Hard	Hidden	Overall
Mask R-CNN	ResNet-101-FPN	83.6	83.6	66.2	77.8
Cascade Mask R-CNN	ResNet-101-FPN	82.5	81.6	66.5	76.9
SDANet	ResNet-101-FPN	84.1	82.7	66.6	77.8
YOLOv5	CSP Darknet53	85.1	88.3	59.0	77.5
RetinaNet	ResNet50	77.8	79.7	57.6	71.7
DAMN	ResNet-101-FPN	86.7	83.7	67.6	79.3



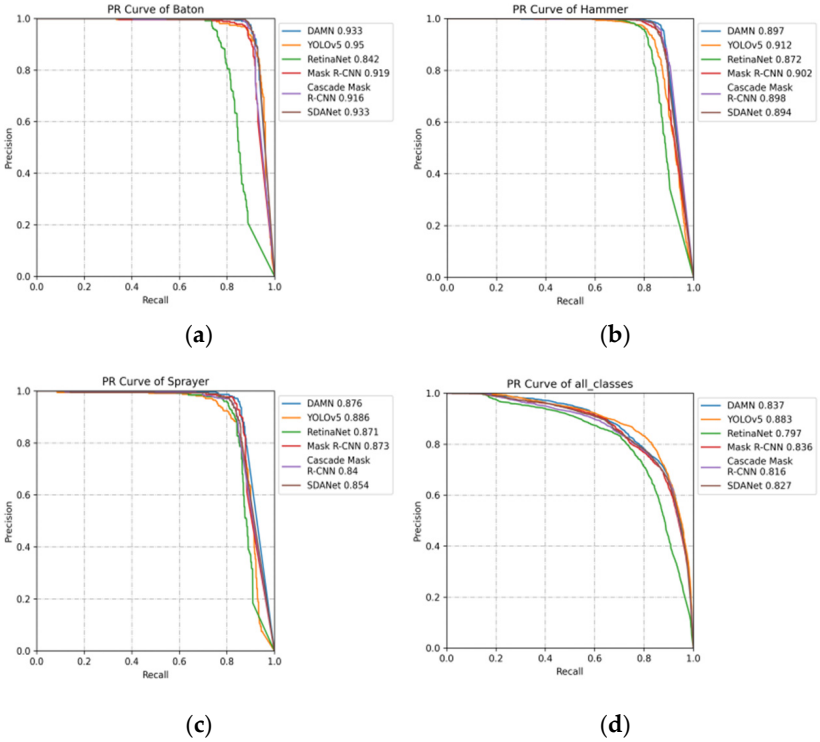
(**a**) (**b**)

Figure 6. *Cont.*



(**c**) (**d**)

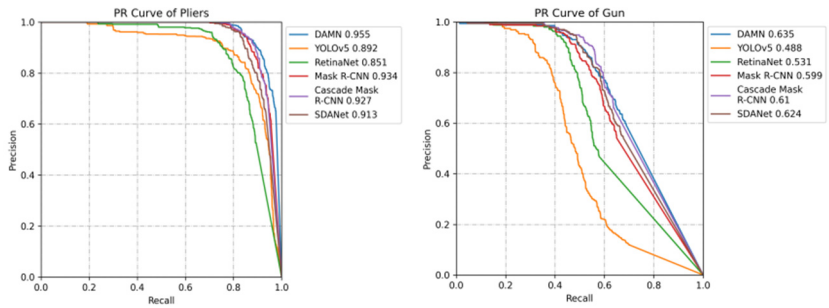
Figure 6. Comparison of PR curves based on easy datasets. (**a**) PR curve of hammer. (**b**) PR curve of power bank. (**c**) PR curve of bullet. (**d**) PR curve of 12 categories.



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

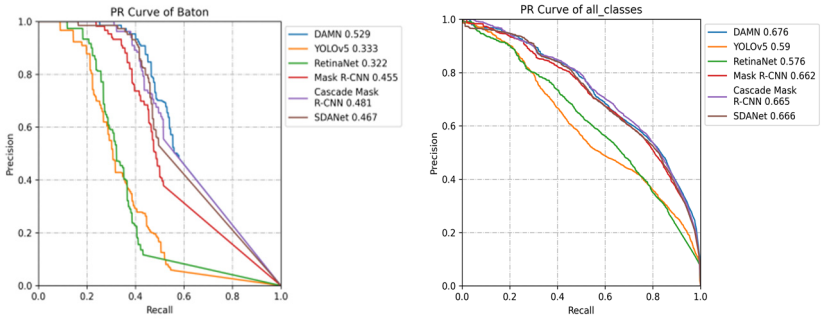
(**c**) (**d**)

Figure 7. Comparison of PR curves based on hard datasets. (**a**) PR curve of baton. (**b**) PR curve of hammer. (**c**) PR curve of sprayer. (**d**) PR curve of 12 categories.



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

Figure 8. *Cont*.



(**c**) (**d**)

Figure 8. Comparison of PR curves based on hidden datasets. (**a**) PR curve of pliers. (**b**) PR curve of gun. (**c**) PR curve of baton. (**d**) PR curve of 12 categories.

On the hidden dataset, the DAMN also achieves the best performance, with an mAP of 67.6%. It is important to note that on the hard dataset, the DAMN achieves an mAP of 83.7%, slightly lower than YOLOv5. This is primarily attributed to YOLOv5's exceptional performance in multiscale detection, enabling it to simultaneously detect objects of various sizes, thus enhancing its overall robustness. However, in the hidden dataset, it is worth noting that the DAMN outperforms YOLOv5 by an 8.6% increase in mAP. This significant advancement can be attributed to the DAMN's incorporation of a dual attention mechanism. The spatial attention mechanism aids in precise target localization, even when objects are occluded or disguised.

Simultaneously, the channel attention mechanism helps capture critical details of the target. The fusion of these spatial and channel attention mechanisms equips the model with a superior understanding of the context and semantics of hidden objects within the image, leading to more accurate detection of concealed targets and reduced false positives. This capability proves particularly advantageous in complex camouflage scenarios. When detecting deliberately concealed items like the sprayer and the hammer, the DAMN continues to excel, especially in scenarios where other models exhibit subpar performance.

In summary, the DAMN consistently demonstrates exceptional performance across the board, achieving an overall mAP of 79.3%. This represents a noteworthy 2.96% average improvement over the other five algorithms. These results underscore the remarkable versatility and robustness of the DAMN, thanks to its integrated dual attention mechanism in the context of X-ray contraband detection tasks.

4.4. Special Case Analysis

4.4.1. Detection of Small Prohibited Items

Figure 9 illustrates a comparative analysis of small object prohibited item detection. In the graph, it is evident that the DAMN excels in detecting small prohibited items with less prominent features and smaller sizes, such as lighters. In contrast to the potential instances of missed detections by other methods, the DAMN consistently maintains high accuracy and effectively identifies the precise location of items like lighters. This accomplishment can be attributed to the multilayer perceptron (MLP) embedded within the spatial attention mechanism, which demonstrates exceptional performance in extracting implicit texture information. Leveraging its powerful spatial attention mechanism, the DAMN successfully captures the precise positions of concealed prohibited items like lighters, achieving an impressive detection accuracy of 0.99.

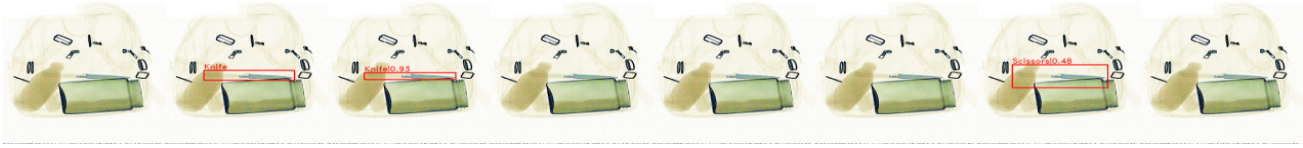
(**a**) Image (**b**) GT (**c**) DAMN (**d**) SDANet (**e**) Cascade MaskR-CNN (**f**) Mask R-CNN (**g**) YOLOv5 (**h**) RetinaNet

Figure 9. Comparison of small prohibited item detection.

This not only highlights the DAMN's outstanding performance in detecting small prohibited items but also demonstrates its robust detection capability when dealing with inconspicuous objects. Such results once again strongly reaffirm the DAMN's superior performance in X-ray prohibited item detection tasks.

4.4.2. Detection of Prohibited Items with Easily Lost Details

For objects composed of multiple materials, such as a small knife made of both metal and wood, the detailed features often tend to be lost. In this context, as shown in Figure 10, except for the DAMN and YOLOv5, other methods exhibit instances of missed detection. It is worth noting that although YOLOv5 wrongly identifies the small knife as a pair of scissors during the detection process, the DAMN not only accurately detects the small knife but also achieves a significant detection accuracy of 0.93. This outstanding performance is attributed to the critical role of spatial attention, which effectively connects features across different spatial locations. Particularly noteworthy is the fact that the spatial attention mechanism enhances the relevant channels responsible for detecting materials like metal and wood, enabling the DAMN to pinpoint the position of the small knife with high precision. This enhancement effect is demonstrated in Figure 10, where even the subtle features of the small knife's handle are clearly discerned by the DAMN.



(**a**) Image (**b**) GT (**c**) DAMN (**d**) SDANet (**e**) Cascade MaskR-CNN (**f**) Mask R-CNN (**g**) YOLOv5 (**h**) RetinaNet

Figure 10. Comparison of detection of prohibited items with easily lost details.

This vividly demonstrates the DAMN's significant breakthrough in detecting details of objects composed of multiple materials. Not only does the DAMN exhibit high accuracy, but it also captures highly challenging subtle features. This further validates the DAMN's superior performance in detecting prohibited items with easily lost details.

4.4.3. Detection of Overlapping and Occluded Prohibited Items

In scenarios involving overlapping and occluded prohibited items, as shown in Figure 11, we can observe the overlapping of the prohibited items, and it is difficult to distinguish the "knife" from the metallic occluder. In such cases, the knife is mis-detected using SDANet, YOLOv5, and RetinaNet. It is even mis-detected as a bullet using Cascade Mask R-CNN. However, the knife is detected successfully using the DAMN due to the collaboration of spatial attention and channel attention.

(**a**) Image (**b**) GT (**c**) DAMN (**d**) SDANet (**e**) Cascade MaskR-CNN (**f**) Mask R-CNN (**g**) YOLOv5 (**h**) RetinaNet

Figure 11. Comparison of detection of overlapping and occluded prohibited items.

In conclusion, spatial attention focuses on extracting semantic and textural information from objects, while channel attention enhances these crucial features and establishes their interconnections. It is the synergistic collaboration of these two attention mechanisms that form the foundation of the DAMN's outstanding capability in detecting prohibited items under complex circumstances.

5. Conclusions

X-ray prohibited item detection technology plays a crucial role in ensuring public safety. However, the significant workload has made automated security screening a current research focus. To address this, this paper introduces an innovative X-ray prohibited item detection method referred to as the DAMN, which integrates spatial attention, channel attention, and dependency optimization modules to significantly enhance detection performance. Specifically, the DAMN offers the following advantages: Firstly, the DAMN makes full use of the spatial attention mechanism, enabling the detector to focus more on the semantic and textural information from various regions of the image, which enhances sensitivity and discrimination. Secondly, the channel attention in the DAMN enhances the representation of crucial features, thereby improving the detector's accuracy and robustness and helping mitigate the impact of interference and noise. Thirdly, the DAMN incorporates a dependency optimization module, which effectively express the contextual relationships among objects, which can improve the detector's capability to handle complex scenes, thereby reducing instances of false positives and missed detections.

In our experiment, the DAMN method comprehensively demonstrates its superior performance across four different categories of datasets: easy, hard, hidden, and overall. Whether dealing with small objects, details prone to loss, or situations involving overlap and occlusion, the DAMN consistently achieves optimal detection accuracy. When compared to mainstream detection algorithms, such as YOLOv5, RetinaNet, Mask R-CNN, Cascade Mask R-CNN, and SDANet, the DAMN outperforms in every aspect, showcasing its remarkable capabilities in the field of X-ray prohibited item detection.

In our future research, we will further incorporate multimodal data fusion with information from other sensors, such as visible light images, infrared images, radar signals, and more, to enhance detection accuracy and robustness, thereby expanding the horizon for the development of X-ray prohibited item detection technology. To protect sensitive data, we will take measures to ensure the privacy and legal compliance of information, including removing identifiable information and adhering to relevant regulations, such as the HIPAA, among others.

Author Contributions: Data curation, C.Z. and S.S.; funding acquisition, Y.L. and G.Y.; investigation, S.S.; methodology, Y.L.; supervision, G.Y.; validation, C.Z. and Y.L.; writing original draft, Y.L.; writing—review and editing, G.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of the Fujian Province under Grant 2021J01865 and the Education and Scientific Research Project for Middle-Aged and Young Teachers of the Fujian Province (No. JAT210678 and No. JT180877) and Fujian Province Undergraduate Higher Education Teaching Research Project (No. FBJY20230164).

Data Availability Statement: We include a data availability statement with all research articles published in an MDPI journal.

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

References

1. Michel, S.; Koller, S.M.; de Ruiter, J.C.; Moerland, R.; Hogervorst, M.; Schwaninger, A. 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; pp. 201–206.
2. Li, Y.; Sun, S.; Zhang, C.; Yang, G.; Ye, Q. One-stage disease detection method for maize leaf based on multi-scale feature fusion. *Appl. Sci.* **2022**, *12*, 7960. [CrossRef]
3. Kundegorski, M.E.; Akcay, S.; Devereux, M.; Mouton, A.; Breckon, T.P. On using feature descriptors as visual words for object detection within X-ray baggage security screening. In Proceedings of the International Conference on Imaging for Crime Detection & Prevention, Madrid, Spain, 23–25 November 2016; pp. 1–6.
4. Akcay, S.; Kundegorski, M.E.; Devereux, M.; Breckon, T.P. 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; pp. 1057–1061.
5. Mery, D.; Svec, E.; Arias, M.; Riffo, V.; Banerjee, S. Modern computer vision techniques for X-ray testing in baggage inspection. *IEEE Trans. Syst. Man Cybern. Syst.* **2016**, *47*, 682–692. [CrossRef]
6. Akcay, S.; Breckon, T.P. An evaluation of region based object detection strategies within X-ray baggage security imagery. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017.
7. Akcay, S.; Kundegorski, M.E.; Willcocks, C.G.; Breckon, T.P. Using deep convolutional neural network architectures for object classification and detection within X-ray baggage security imagery. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2203–2215. [CrossRef]
8. Liu, J.; Leng, X.; Liu, Y. Deep convolutional neural network based object detector for X-ray baggage security imagery. In Proceedings of the IEEE International Conference on Tools with Artificial Intelligence, Portland, OR, USA, 4–6 November 2019.
9. Bhowmik, N.; Wang, Q.; Gaus, Y.F.A.; Szarek, M.; Breckon, T.P. The good, the bad and the ugly: Evaluating convolutional neural networks for prohibited item detection using real and synthetically composited X-ray imagery. *arXiv* **2019**, arXiv:1909.11508.
10. Subramani, M.; Rajaduari, K.; Choudhury, S.D.; Topkar, A.; Ponnusamy, V. Evaluating one stage detector architecture of convolutional neural network for threat object detection using X-ray baggage security imaging. *Rev. d'Intelligence Artif.* **2020**, *34*, 495–500. [CrossRef]
11. Su, Z.G.; Yao, S.Q. A multi-object prohibited items identification algorithm based on semantic segmentation. *J. Signal Process.* **2020**, *36*, 7.
12. Gu, J. A Study on X-ray Safety Check Contraband Image Detection Based on Deep LEARNING. Master's Thesis, Yunnan University, Kunming, China, 2021.
13. Dong, Y.S.; Li, Z.X.; Guo, J.Y.; Chen, T.Y.; Lu, S.H. Improved YOLOv5 model for X-ray prohibited item detection. *Adv. Lasers Optoelectron.* **2023**, *60*, 8.
14. Li, S.; Ya, S.; Mu, S. Improved YOLOv7 X-ray image real-time detection of prohibited items. *Comput. Eng. Appl.* **2023**, *59*, 193–200.
15. Han, N. *A Deep Learning-Based Dangerous Goods Detection and Tracking Algorithm from X-ray Images*; Lanzhou University: Lanzhou, China, 2018.
16. Zhang, Y.K.; Su, Z.G.; Zhang, H.G.; Yang, J.F. Multi-scale prohibited item detection in X-ray security image. *J. Signal Process.* **2020**, *36*, 11.
17. Zhang, Z.; Li, M.Z.; Li, H.F.; Ma, J.Q. Improved SSD algorithm and its application in subway security detection. *Comput. Eng.* **2021**, *47*, 7.
18. Ren, J. *X-ray Security Inspection Image Contraband Detection Based on YOLOv5*; China University of Geosciences (Beijing): Beijing, China, 2021.
19. Mu, S.Q.; Lin, J.J.; Wang, H.Q.; Wei, X.Z. An Algorithm for detection of prohibited items in X-ray images based on improved YOLOv4. *Acta Armamentarii* **2021**, *42*, 2675–2683.
20. Kang, J.N.; Zhang, L. Multi-scale X-ray security inspection image detection with multi-channel region proposal. *Comput. Eng. Appl.* **2022**, *58*, 8.
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. *arXiv* **2016**, arXiv:1612.03144.
22. Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer: Berlin/Heidelberg, Germany, 2006.
23. Wu, Z.; Gobichettipalayam, S.; Tamadazte, B.; Allibert, G.; Paudel, D.P.; Demonceaux, C. Robust rgb-d fusion for saliency detection. In Proceedings of the 2022 International Conference on 3D Vision (3DV), Prague, Czechia, 12–15 September 2022; pp. 403–413.
24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
25. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–4 September 2018; pp. 3–19.
26. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.

27. Qu, Y.; Shao, Z.; Qi, H. Non-Local Representation based Mutual Affine-Transfer Network for Photorealistic Stylization. *arXiv* **2019**, arXiv:1907.10274. [CrossRef]
28. Wang, B.; Zhang, L.; Wen, L.; Liu, X.; Wu, Y. Towards real-world prohibited item detection: A large scale X-ray benchmark. *arXiv* **2021**, arXiv:2108.07020.
29. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Volume 13, pp. 740–755.
30. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1483–1498. [CrossRef] [PubMed]