electronics

MDPI

Article

Efficient X-ray Security Images for Dangerous Goods Detection Based on Improved YOLOv7

Yan Liu [1], Enyan Zhang [2], Xiaoyu Yu [1,*] and Aili Wang [2]

[1] College of Electron and Information, University of Electronic Science and Technology of China, Zhongshan Institute, Zhongshan 528402, China; yanliu@zsc.edu.cn
[2] Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application, Harbin University of Science and Technology, Harbin 150080, China; 2220610136@stu.hrbust.edu.cn (E.Z.); aili925@hrbust.edu.cn (A.W.)
* Correspondence: yuxy@zsc.edu.cn

Abstract: In response to the problems of complex background, multi-scale dangerous goods and severe stacking in X-ray security images, this paper proposes a high-accuracy dangerous goods detection algorithm for X-ray security images based on the improvement of YOLOv7. Firstly, by combining the coordinate attention mechanism, the downsampling structure of the backbone network is improved to enhance the model's target feature localization ability. Secondly, a weighted bidirectional feature pyramid network is used as the feature fusion structure to achieve multi-scale feature weighted fusion and further simplify the network. Then, combined with dynamic snake convolution, a downsampling structure was designed to facilitate the extraction of features at different scales, providing richer feature representations. Finally, drawing inspiration from the idea of group convolution and combining it with Conv2Former, a feature extraction module called a multi-convolution transformer (MCT) was designed to enhance the network's feature extraction ability by combining multi-scale information. The improved YOLOv7 in this article was tested on the public datasets SIXRay, CLCXray, and PIDray. The average detection accuracy (mAP) of the improved model was 96.3%, 79.3%, and 84.7%, respectively, which was 4.7%, 2.7%, and 3.1% higher than YOLOv7. This proves the effectiveness and universality of the method proposed in this article. Compared to the current mainstream X-ray image dangerous goods detection models, this model effectively reduces the false detection rate of dangerous goods in X-ray security inspection images and has achieved significant improvement in the detection of small and multi-scale targets, achieving higher accuracy in dangerous goods detection.

Keywords: X-ray images; dangerous goods detection; attention mechanism; dynamic convolution; transformer

Citation: Liu, Y.; Zhang, E.; Yu, X.; Wang, A. Efficient X-ray Security Images for Dangerous Goods Detection Based on Improved YOLOv7. *Electronics* **2024**, *13*, 1530. https://doi.org/10.3390/electronics13081530

Academic Editor: Young-Ho Seo

Received: 25 March 2024
Revised: 13 April 2024
Accepted: 16 April 2024
Published: 17 April 2024

1. Introduction

The installation of X-ray security checks at the entrance of important transportation hubs aims to maintain the safety and order of public places, which is conducive to ensuring the personal and property safety of citizens [1]. The absorption and scattering attenuation of X-rays vary among different materials, and different items appear in different colors in X-ray images. Currently, X-ray security equipment is widely used due to its convenience and high efficiency in completing prohibited item detection tasks without opening the luggage of the inspected personnel. However, it requires personnel to visually locate and inspect the prohibited items in the X-ray images generated by the security equipment. X-ray images have the characteristics of complex backgrounds, diverse targets, low color contrast, and severe occlusion. Long-term work by security personnel can lead to false or missed detections and even major safety accidents due to fatigue [2].

The task of detecting dangerous goods in X-ray security images can be roughly divided into two parts: locating and classifying dangerous goods in luggage. Specifically, it uses

object detection algorithms to confirm whether the detected X-ray image of the luggage contains dangerous goods, such as knives and guns, and provides feedback on the location and category information of the target. Compared to commercial systems supplied with existing luggage screen devices, our method has higher accuracy in detecting dangerous goods and a lower probability of missed or false detections. The characteristics of X-ray security inspection images are summarized as follows:

(1) Background complexity: Due to the different transmittance of X-rays through different objects, objects with different materials and thicknesses exhibit different colors under X-rays. Objects with similar thicknesses to hazardous materials can interfere with the model's learning of hazardous material feature information;
(2) Serious overlap phenomenon: The overlapping stacking of items can obscure the edges of objects, reducing the characteristic of dangerous goods and increasing the difficulty of dangerous goods detection;
(3) Multi-scale dangerous goods: In X-ray images, dangerous goods, such as knives and liquid containers, have multiple categories, shapes, and scales. Different types of dangerous goods have different sizes, and there are also differences in the size and shape of the same category of dangerous goods. Dangerous goods have intra-class and inter-class differences. Therefore, complex backgrounds, overlapping occlusion of hazardous materials, and multiple scales are the main problems faced by current X-ray security image hazardous material detection tasks.

In 2011, Bastan introduced the visual word bag model into the field of X-ray security, compared common feature representation methods, and found that the strategy performance of combining Gauss and scale-invariant feature transform (SIFT) feature descriptors was the most outstanding, with a mean average precision (mAP) of 65% [3]. In 2013, Tursany proposed a new bag of words representation model for gun detection tasks, which used feature clustering to separate features of different categories for detection targets. The combination of a speeded up robust features (SURF) feature descriptor and a support vector machine (SVM) classifier achieved significant improvement in the classification performance of the model [4]. In 2015, Fliton et al. compared the performance of different feature descriptors combined with classification algorithms for pistols and bottles in 3D CT images and found that the density histogram-based feature descriptors DH (density histogram) and DGH (density gradient histogram) performed better in classification tasks than SIFT, rotation-invariant feature transform (RIFT), and other feature description methods [5]. In 2017, Mery et al. proposed an X-ray imaging model for detecting targets, which uses an adaptive sparse representation method to separate targets and backgrounds, improving the classification performance of hazardous material detection in X-ray images. Liu Ying et al. proposed an automatic detection algorithm for X-ray security inspection images with a biased color texture dictionary. The algorithm uses a color weight matrix to extract local texture features of the image and combines it with a word bag model to optimize multiple channels to improve detection accuracy [6]. In 2018, Xing Xiaolan et al. proposed a grayscale projection algorithm based on security inspection images, which used median filtering to denoise the images, combined with a grayscale integral projection algorithm and block grayscale projection algorithm to identify suspected hazardous material areas, and improved processing speed through FPGA deployment [7]. Russo et al. used an approximate median filtering algorithm to separate the target from the background, then used a target contour-based filtering method to locate the region of interest and generate feature vectors, and finally used SVM to classify the target [8]. In 2019, Lyu et al. proposed a high-speed X-ray image classification algorithm for sorting packages, dividing X-ray image samples into three categories: normal, dangerous, and suspicious. By extracting the feature information of dangerous goods through affine moment invariants, normal and dangerous samples can be automatically processed, requiring only manual processing of the samples, saving a lot of human resources [9]. The X-ray security inspection image hazardous material detection method based on traditional methods has the characteristic of clear logical theory. Although it has strong interpretability, feature extraction overly relies

on manual labor, resulting in low efficiency and poor detection performance in complex background detection tasks without generalization.

In recent years, research on object detection based on deep learning has reached a new peak. In 2016, Akcay et al. applied convolutional neural networks (CNNs) for the first time to X-ray luggage image classification tasks and detected dangerous goods contained within them. Through a series of experiments, it was demonstrated that CNN-related technologies have broad prospects in X-ray security images [10]. In 2017, Mery et al. tested over ten security image recognition methods on the GDXray dataset, including those based on Bow, sparse KNN, deep learning, etc. [11]. In 2018, Singh et al. proposed the R-FCN3000 model to optimize faster RCNN in response to the problem of duplicate ROI calculation during the training process. The model decouples detection and classification, improving detection speed while also having strong generalization and accuracy [12]. Morris et al. proposed hazardous material detection based on VGG19, Xception, and InceptionV3 models. The ROC curve of the optimal model was 0.95 [13]. Gaus et al. proposed a dual convolutional neural network architecture for automatic anomaly detection in complex and secure X-ray images, which utilizes region-based R-CNN and mask-based R-CNN to provide localization variants for target categories of interest, optimizing intra-object anomaly detection into two types of problems: anomaly or benign [14]. However, most of the above object detection methods are two-stage detection algorithms based on anchor boxes. Although the detection accuracy of dangerous goods in security inspection images is high, it is necessary to first generate candidate boxes corresponding to the target and then conduct distribution training to share the detection results. This makes the detection speed reach the neck of the bottle, which cannot meet the real-time requirements in security inspection scenes.

In 2021, Da Huang Li proposed a new method based on generative adversarial networks (GANs) to synthesize X-ray safety images with multiple prohibited items from semantic label images. The dataset was expanded by synthesizing a limited number of samples, and single-shot detector (SSD) was used as the experimental detection algorithm with an mAP of 82.50% [15]. In response to the issue of scale differences in prohibited item detection, Wang Yuxiao et al. proposed a multi-scale feature fusion detection network (MFFNet) based on an improved SSD [16]. The deep RestNet-101 network was used as the backbone network of the SSD to enhance the network's feature extraction ability. In addition, a lightweight feature fusion module was used to generate a new feature pyramid, and the improved model was tested on the SIXray dataset. The detection accuracy has significantly improved, but the detection speed has decreased by nearly three times [17]. Tang Haoyang et al. improved the SSD detection algorithm using a feature pyramid structure to fuse the shallow and deep positional and semantic information of the feature map. In addition, they introduced a deformable convolution module to automatically adjust and obtain the contour and scale information of the detection target, enhancing the detection ability for small targets [18]. Zhang Youkang et al. proposed an asymmetric convolutional multi-view detection network, which uses small convolutional asymmetric networks, multi-scale feature map fusion strategies, and dilated multi-view convolution modules to improve the identification of dangerous goods in X-ray security inspection images under background interference [19]. In order to solve the problem of overlapping dangerous goods, Wei et al. proposed a de-occlusion attention module (DOAM), which considers the edge information and material information of dangerous goods in X-ray images from the perspective of attention, thereby improving detection performance [20]. In order to solve the problem that some dangerous goods detection methods cannot guarantee good performance when applied to multiple datasets, Yang et al. proposed a new dual-mode learning network (DML-Net) to effectively detect dangerous goods in multiple datasets [21].

Lu Guanyou et al. improved the YOLOv3 network model by using the K-means clustering algorithm to calculate prior boxes that fit the target size and improved detection speed by reducing the predicted bounding boxes [22]. Wu Haibin et al. optimized the YOLOv4 network structure using dilated convolution and introduced a spatial pyramid pooling model to increase the receptive field and enhance the detection ability of small

target dangerous goods. They demonstrated a detection performance of 85.23% mAP on the SIXray dataset [23]. Dong Yishan et al. proposed an improved YOLOv5 network model, which introduces a convolutional block attention module to enhance the network's feature extraction ability for targets [24]. Zhang et al. proposed an X-ray dangerous goods detection network, ScanGuard YOLO, based on YOLOv5. Firstly, dilated convolution was used to increase the receptive field of the backbone network. Secondly, an efficient multi-scale feature fusion module, RepGFPN, was designed to fuse multi-scale information, effectively improving the recall rate and F1 score of the model for detecting hazardous materials [25]. Li et al. proposed a large-scale X-ray image dataset LSIray for dangerous goods detection and proposed an improved detection method, SC-YOLOv8. By using deformable convolution to adapt to the diversity of targets and using a spatial pyramid multi-head attention mechanism (SPMA) to enhance the representation ability of targets, the LSIray dataset showed a detection performance of 82.7% mAP [26].

The one-stage object detection method represented by the YOLO series mentioned above does not require generating candidate regions and directly predicts the category and location of hazardous materials based on regression methods, which can ensure high detection accuracy and greatly improve detection speed.

In response to the issues of false positives and missed detections caused by the characteristics of X-ray security images, Zhang Hong et al. proposed an adaptive feature fusion strategy by improving the spatial and channel attention mechanisms, making the learning of the YOLOv5 network more targeted [27]. Han Ping et al. proposed a regional enhancement and multi-feature fusion model called REMF to adapt to detecting chaotic and complex items in images, and they optimized the fusion effect using a ternary loss function [28]. On the basis of YOLOv7, Cheng Lang et al. improved efficient long-range attention by adding skip connections and 1×1 convolutional branches between the constituent modules, enhancing the network's feature extraction ability while accelerating inference speed. For the problem of ignoring target directionality in prohibited item detection, the method of dense set encoding labels was used to discretize the angle, which improved the accuracy of prohibited item localization [29].

The above methods have improved the accuracy of dangerous goods detection in X-ray security images, providing diverse ideas for the application of deep learning in the field of intelligent security. However, in response to the issue of intra- and inter-class differences in dangerous goods, existing sampling methods with fixed positions cannot combine multi-scale feature information for prediction. In addition, X-rays have transmissivity, and items of the same material have similar feature information when transmitted through X-rays, blurring the boundary between the target and background and making it difficult to distinguish, causing interference in recognition.

At present, large-scale language models are becoming increasingly widely used due to their powerful ability to handle complex datasets. Large language models (LLMs), such as LLaMA-2, can dynamically adjust and interpret various instructions in embedded languages. When providing real-time X-ray security images, the combination of multimodal LLaMA-2 and image hashing ensures real-time detection efficiency through the offline construction of a description library. It can accurately locate dangerous goods, potential obstructions, and even fine details such as the construction materials of dangerous goods in luggage [30]. When writing text to call large language models such as ChatGPT and GPT-4, users are required to use chat interfaces or application programming interfaces to send data to OpenAI for processing. Considering the privacy of security data, the large language model Vicuna-13B can be chosen to run locally in a privacy-preserving manner. This model fine-tunes LLaMA-2 in a supervised manner, with fewer parameters compared to ChatGPT and GPT-4 and performance comparable to ChatGPT [31]. The YOLO series detectors have a strong dependence on the category of pre-trained detection targets. To address this limitation, Cheng et al. proposed YOLO-World [32], which utilizes a new network called RepVL-PAN (Re-parametric Vision-Language Path Aggregation Network) to facilitate the interaction between visual and linguistic information. By using open

vocabulary for detection, the detector's applicability in open scenes is enhanced. Matthias Minderer et al. used a recipe called OWL-ST [33] to enrich weakly supervised web data, overcoming dependence on human annotations by expanding the scale of self-training and breaking free from the limitations of manual annotation detection training data on open vocabulary detection performance.

The current field of object detection mostly revolves around YOLOv7. The related technologies of YOLOv7 are relatively mature, and it is convenient to compare and improve using YOLOv7. Moreover, in recent years, the iteration speed of YOLO versions has been fast, and the difference between the two versions is not significant. We propose an improved YOLOv7 network that combines a coordinate attention mechanism, snake-shaped dynamic convolution, and Conv2Former to implement a dangerous goods detection algorithm for X-ray security inspection images. The main contributions of the proposed method are summarized as follows:

1. In the downsampling part of the backbone network, coordinate attention is combined to focus on the spatial position information of the input image, making the deep convolutions in the network more sensitive to the position information of the feature map and effectively improving the detection ability of the model. In addition, BiFPN (Bidirectional Feature Pyramid Network) is used as a feature fusion structure to simplify the network and enhance its feature fusion capability.
2. In the downsampling part of the neck network, dynamic snake-shaped convolution is combined as a cross-scale embedding layer for feature extraction at different scales to enhance the robustness of the detection model to changes in the shape and position of dangerous goods, and multi-scale feature information is considered to improve the localization and recognition ability of dangerous goods in complex backgrounds.
3. To design a multi-scale grouped convolution module, a multi-convolution transformer (MCT) block based on Conv2Former was used to simplify the self-attention mechanism in the vision transformer and further refine the key feature layers to prevent the loss of feature information at different scales. This effectively reduces computational costs while processing high-resolution images and enhancing the network's ability to extract local feature information of hazardous material.

2. Methods

2.1. Overview

As a representative of the first-stage object detection algorithm, YOLO's core idea is to treat the object detection problem as a regression problem by dividing the image into several grids and predicting the bounding boxes of each cell, fully extracting the features of the detected target. Considering that the application scenario is the detection of prohibited items in important transportation hubs, this requires the detection model to balance accuracy and real-time performance. YOLOv7, as a relatively advanced first-stage detection algorithm, demonstrates excellent detection capabilities and is suitable as a basic model for dangerous goods detection.

In the X-ray security inspection of dangerous goods detection tasks, YOLOv7 uses special convolutional and pooling layers as downsampling modules to extract advanced feature information related to dangerous goods. In addition, it can increase the receptive field and improve the network's detection ability for dangerous goods. However, as the size of the stacked feature maps of the convolutional layers decreases, the resolution becomes lower, resulting in the loss of some positional information in the feature maps. This can lead to a decrease in the accuracy of locating dangerous goods in the background of highly overlapping items in luggage, which is very unfavorable for object detection tasks, especially small object detection tasks.

The improved YOLOv7 network structure is shown in Figure 1.

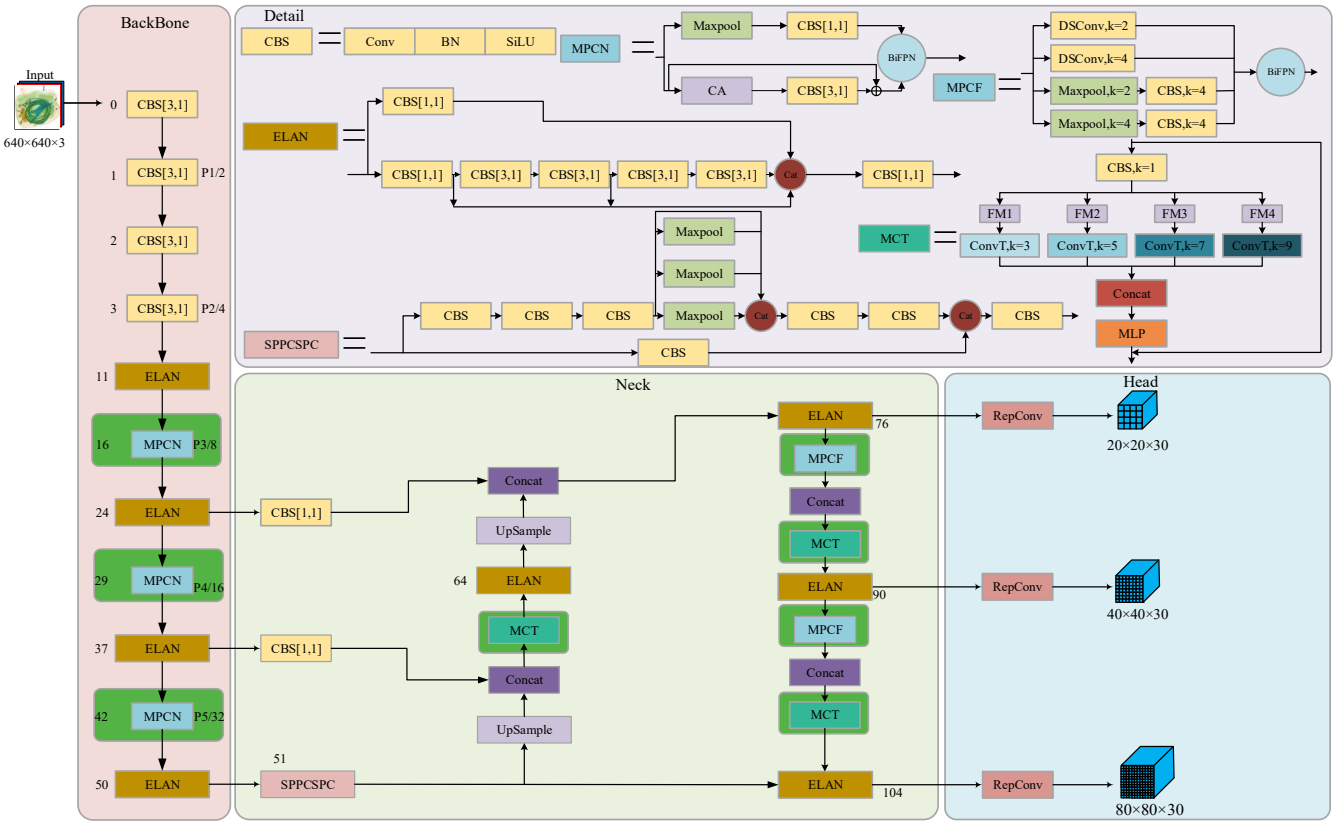


Figure 1. The network structure of improved YOLOv7.

To address the aforementioned issues, improvements were made to the MaxPool and convolution (MPConv) structures in the YOLOv7 backbone network and neck network to enhance the robustness of the model to changes in the size profile of hazardous materials. In addition, there is a problem of scale differences in hazardous material detection tasks, as hazardous materials have multi-scale characteristics. There are differences in the size, contour, and other characteristic information of different types of hazardous materials in the same X-ray security image. The relative size of hazardous materials in the same category also varies in different X-ray security images. Multi-scale hazardous material detection is another difficulty. To address this issue, the MCT module combined with Conv2Former and multi-scale feature map information are used to enhance the network's feature extraction capability.

2.2. Improvement of MPConv Module in Backbone Network

During the downsampling process of the YOLOv7 object detection model, as the size of the stacked feature maps of the convolutional layers decreases, the resolution becomes lower, resulting in the loss of some positional information in the feature maps, which can lead to a decrease in the accuracy of locating prohibited items in the background of highly overlapping items in luggage. Therefore, by combining the coordinate attention mechanism and weighted bidirectional feature pyramid network, the MPConv module of the backbone part is improved to obtain the MPCN module. MPCN replaces the 3×3 ordinary convolution with the CA (coordinate attention) module to extract local small-domain feature information, which focuses on the position information of different targets in X-ray security images, avoiding the interference of a large amount of feature information of objects similar to dangerous goods in the detected package on model training. Secondly, BiFPN is used as the feature fusion structure, introducing cross-scale connections and multi-scale weighted fusion to reduce the computational and parameter complexity of the model. Finally, referring to the residual network structure in the RepVgg architecture, a

skip connection is added to the lower branch to solve the problem of gradient dispersion or vanishing in the main branch where the CA module is located. The MPCN structure is shown in Figure 2.

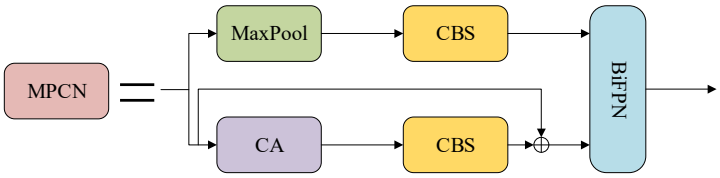


Figure 2. The structure of MPCN.

2.2.1. Coordinate Attention Mechanism

In the backbone network, the ordinary convolution size of 3 × 3 used to extract local small-area feature information will be replaced with the CA (coordinate attention) module, which focuses on the position information of different targets in X-ray security images to avoid interference from the large amount of feature information of objects similar to dangerous goods in the detected package on model training. The structure of CA is shown in Figure 3.

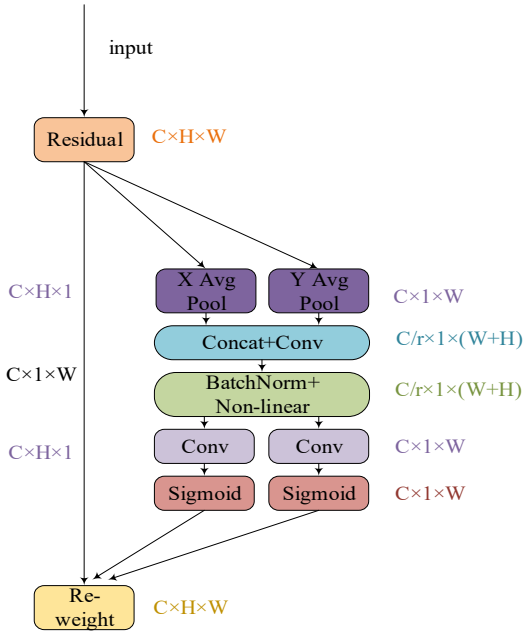


Figure 3. The structure of coordinate attention.

The coordinate attention mechanism includes embedding coordinate information and generating coordinate attention to obtain cross-channel and position information. In the process of embedding coordinate information, the size of the input feature map X is first set as $C \times H \times W$, where C is the number of channels, H is the height, and W is the width. Firstly, a convolutional pooling kernel with size $H \times 1$ and $1 \times W$ is used to decompose the input image feature map into two one-dimensional encoding processes along the horizontal and vertical directions. X_c is the feature of X on the c-th channel. Let the directional aware feature with height h and width w on the c-th channel be output as $Z_c^h(h)$ and $Z_c^w(w)$, and i and j represent the coordinate values.

$$z_c^h(h) = \frac{1}{W_0} \sum_{0 \le i \le W} x_c(h, i) \tag{1}$$

$$z_c^w(w) = \frac{1}{H_0} \sum_{0 \le j \le H} x_c(j, w) \tag{2}$$

In the second stage of the CA module, which is the coordinate attention generation process, the channels in the previous stage are first aggregated along the horizontal and vertical spatial directions to obtain feature maps z^h and z^w, followed by channel compression. After batch normalization and activation function, an intermediate feature map $f \in R^{C/r \times 1 \times (W+H)}$ containing two directions is obtained, where r is the compression ratio.

2.2.2. Weighted Bidirectional Feature Pyramid Network

By using BiFPN as the feature fusion structure, this structure is different from PANet because it not only has its bidirectional feature fusion path but also adopts a cross-scale connection method. The PANet and BiFPN structures are shown in Figure 4.

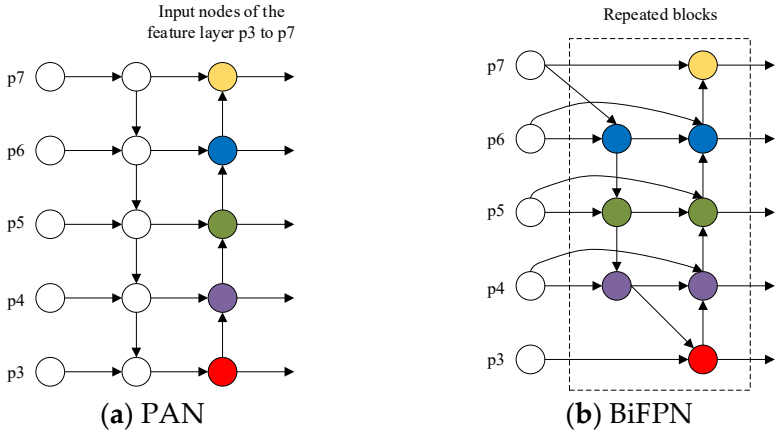


Figure 4. The structure of PAN and BiFPN.

The input feature layer is represented by p_3 to p_7; then, multi-scale weighted fusion and a series of cross-scale-based operations are used in the BiFPN structure to output the extracted features. The simplification of feature fusion networks by BiFPN is reflected in the removal of nodes with only one input, as these single-input nodes do not have feature fusion. Add a branch between the input node and the output node to simplify the network structure and enhance the fusion of more features in the feature fusion network. Introducing cross-scale connections and multi-scale weighted fusion reduces the computational and parameter complexity of the model.

2.3. Improvement of MPConv Module in Neck Network

The neck network of YOLOv7 adopts the PANet design of the feature pyramid (FPN) architecture to achieve an efficient fusion of features at different levels. The MPConv module uses max pooling and a 3×3 convolution kernel with a stride of 2 to implement the downsampling process of the feature fusion part, which cannot establish the interaction between features at different scales. Therefore, a downsampling structure, MPCF, was designed that is more conducive to extracting features of different scales. First, a maximum pooling layer of 4×4 was added to obtain a 4-fold downsampling feature map, reducing the loss of key feature information caused by the decrease in feature map resolution. Secondly, drawing inspiration from the architecture of CrossFormer, we use dynamic snake convolution (DSConv) as a cross-scale embedding layer (CEL) to sample the patches of the input image separately, and then connect the corresponding patch projections into an embedding layer. DSConv can make the network pay more attention to local features related to dangerous goods, such as slender and curved features, and enhance its perception ability of key features. The structure of MPCF is shown in Figure 5.

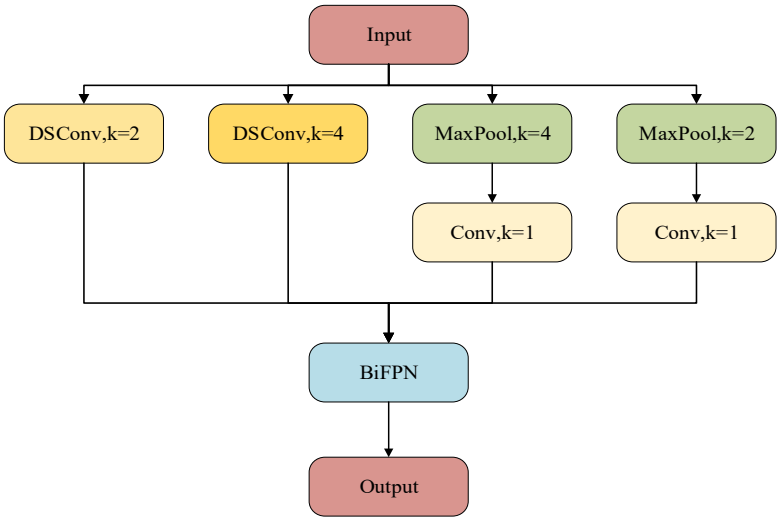


Figure 5. The structure of MPCF.

2.3.1. Dynamic Snake Convolution

In X-ray hazardous material detection tasks, the size and shape of hazardous materials usually vary greatly, the distribution of contextual information is uneven, the sampling position of the standard convolutional rectangular receptive field is fixed, and it cannot be dynamically adjusted according to the target. At the same time, it cannot handle the high variation of contextual information, which leads to the loss of some contextual information and weakens the network's feature extraction ability. Unlike the standard convolutional rectangular receptive field, DSConv makes the receptive field more flexible and adaptable to match various shapes. Dynamic snake convolution learns deformation based on the input feature map to make the deformation conform to the characteristics of tubular structures. The model will pay more attention to local features related to dangerous goods, such as slender and curved features. The convolution kernel is dynamically twisted like a snake to conform to the target structure, thereby achieving more accurate feature extraction. Its structure is shown in Figure 6.

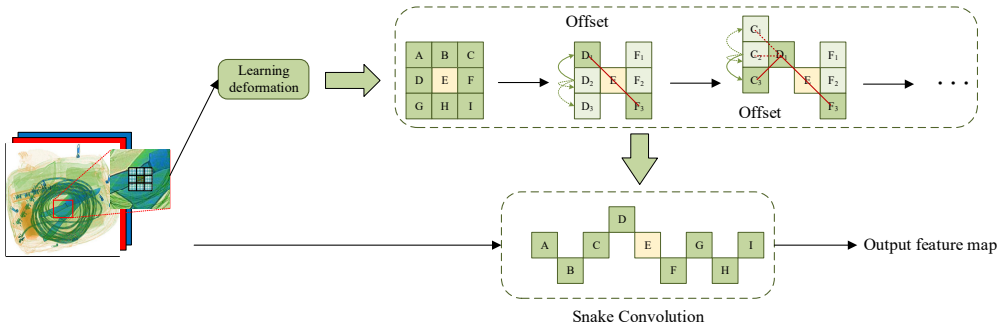


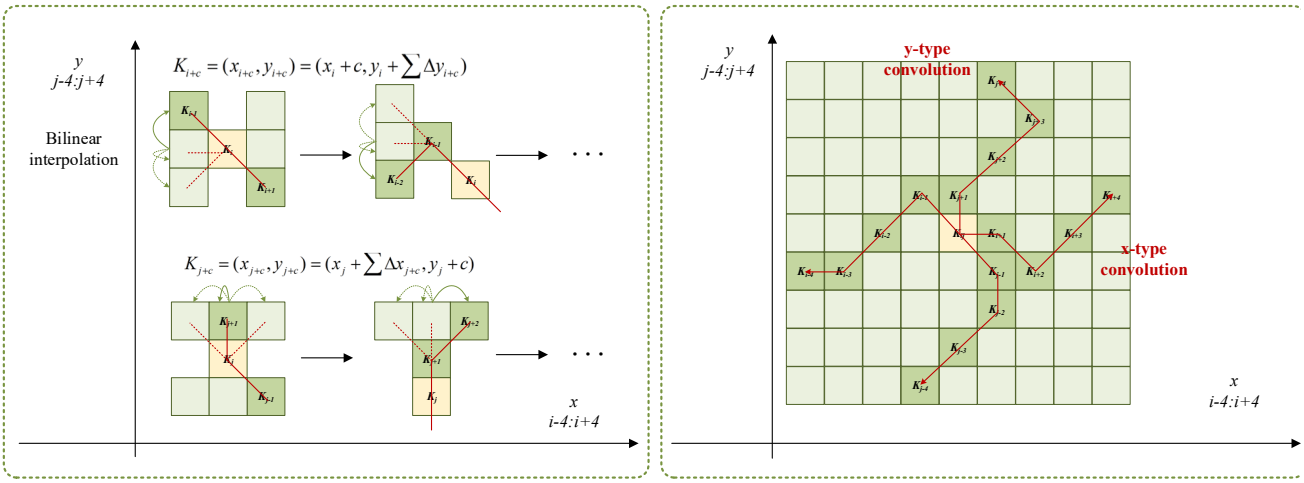
Figure 6. The deformation process of dynamic snake convolution.

The process of extracting local features using dynamic snake-shaped convolution is as follows: assuming the coordinates of the standard 2D convolution are K and the center coordinates are $K_i = (x_i, y_i)$, then a standard convolution with a kernel of 3×3 and coordinates of K can be represented as

$$K = \{(x-1, y-1), (x-1, y), \cdots, (x+1, y+1)\}. \tag{3}$$

Introducing deformation offset Δ makes the convolution kernel more focused on the geometric features of hazardous material targets, but in complex X-ray backgrounds, when the model freely learns deformation offset, the receptive field will deviate from the

target. Therefore, the iterative strategy, as shown in Figure 7, is adopted to select the corresponding observation position for each target to be detected in order to ensure the continuity of attention and not cause the receptive field to propagate too far due to large deformation offsets.



(**a**) Illustration of coordinate caculation of DSConv (**b**) The receptive filed of DSConv

Figure 7. Illustration of DSConv.

In DSConv, the standard convolution kernel is extended along the x-axis and y-axis, with a convolution kernel of 9×9. Taking the x-axis direction as an example, the specific position of each grid in K can be expressed as $K_{i\pm c} = (x_{i\pm c}, y_{i\pm c})$, where $c = \{0, 1, 2, 3, 4\}$ represents the horizontal distance from the central grid. The selection of each grid position $K_{i\pm c}$ in the convolutional kernel K is a cumulative process. Starting from the center position K_i, the position K_{i+1} away from the center grid depends on the position of the previous grid, which adds a deformation offset $\Delta = \{\delta | \delta \in [-1, 1]\}$. Therefore, the offset needs to use $\sum$ to ensure that the convolution kernel conforms to a linear morphological structure, and the offset of coordinates becomes

$$K_{i\pm c} = \begin{cases} (x_{i+c}, y_{i+c}) = (x_i + c, y_i + \sum_{i}^{i+c} \Delta y) \\ (x_{i-c}, y_{i-c}) = (x_i - c, y_i + \sum_{i-c}^{i} \Delta y) \end{cases}. \tag{4}$$

The offset of coordinates changes in the y-axis direction to

$$K_{j\pm c} = \begin{cases} (x_{j+c}, y_{j+c}) = (x_j + \sum_{j}^{j+c} \Delta x, y_j + c) \\ (x_{j-c}, y_{j-c}) = (x_j + \sum_{j-c}^{j} \Delta x, y_i - c) \end{cases}. \tag{5}$$

Considering that offsets Δ are usually fractions, bilinear interpolation can be written as

$$K = \sum_{K'} B(K', K) \cdot K'. \tag{6}$$

Among them, K represents the fractional positions of Equations (4) and (5), and K' lists all integral space positions. B is a bilinear interpolation kernel, which is divided into two one-dimensional kernels:

$$B(K', K) = b(K_x, K'_x) \cdot b(K_y, K'_y). \tag{7}$$

As shown in Figure 7, due to the two-dimensional variation (x-axis, y-axis), DSConv covers a range of 9 × 9 during the deformation process. Its structural design aims to better adapt to the dynamic characteristics of slender tubular structures and enhance its ability to perceive key features.

2.3.2. Cross-Scale Embedding Layer

In order to build cross-scale interaction, drawing on the idea of the cross-scale embedding layer in CrossFormer, DSConvs of different kernel sizes are used to sample the patches of the input image separately, and then the corresponding patches are projected and connected into an embedding layer. As shown in Figure 8, the input image is sampled by convolutions with different kernel sizes and the same step size, and each embedding layer is constructed by projecting and connecting the corresponding patches. Considering that larger kernels can result in higher computational complexity, the principle of "using low dimensionality for larger kernels and high dimensionality for smaller kernels" is adopted to set the projection size for each scale. Compared with the average allocation of projection dimensions, this scheme saves a lot of computational costs but does not affect model performance. The use of DSConv with different kernel sizes as cross-scale embedding layers solves the drawback of each layer's input embedding being equi-scale and unable to extract cross-scale features.

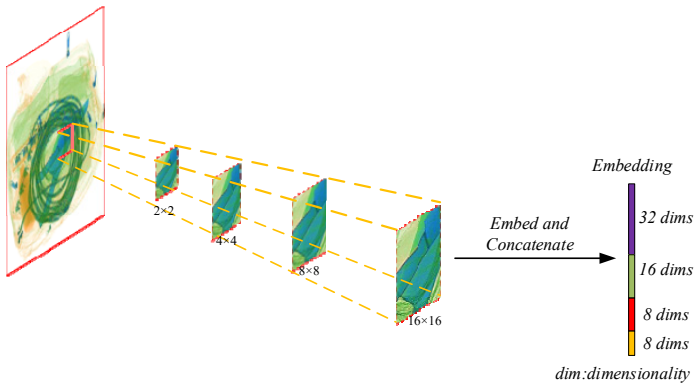


Figure 8. The sampling process of cross-scale embedding layer.

2.4. MCT Module Combined with Conv2Former

Due to their rise in popularity in recent years, vision transformers have been introduced into the YOLO object detection network structure to solve the problem of small object detection. The transformer learns the positional relationship between different features by encoding each feature embedding position, achieving target localization in high-density scenes. However, adding a transformer module to the YOLOv7 network will greatly increase the model's parameter count and computational cost. Therefore, we adopt Conv2Former block modules with different kernel sizes to simplify the self-attention mechanism, extract features from the sub-feature maps of the input feature map separately, and improve the detection accuracy of dangerous goods.

2.4.1. Conv2Former

Conv2Former is a convolutional network visual baseline model built in the transformer style. By using convolutional modulation operations to simplify the self-attention mechanism in the transformer, it can better utilize the nested large kernels ($\geq 7 \times 7$) in the convolutional layer to more effectively encode spatial features using convolution. The traditional attention module uses dot product calculation when calculating similarity as follows:

$$Sim(q_i, k_j) = \frac{q_i \cdot k_j^T}{\sqrt{d_k}}. \quad (8)$$

Among them, q_i and k_j are the query and key vectors in the attention mechanism, respectively, d_k represents the dimension size of the key, and T represents the transpose of the vector. To further simplify the sel- attention mechanism, reduce computational costs, and improve the inference speed of model training, the Hadamard product $\odot$ is used for feature information weighting, which has a normalization effect and is more controllable compared to self-attention mechanism calculation. The similarity calculation formula is written as follows, where $\odot$ is Hadamard product:

$$Sim(q_i, k_j) = \frac{q_i \odot k_j}{\sqrt{d_k}}. \tag{9}$$

The operation process of self-attention and convolutional modulation is shown in Figure 9. Assuming the spatial size of the input is H and W, and the number of channels is C, the self-attention mechanism in the transformer can simulate global pairwise dependencies, providing a more effective spatial encoding method. However, the computational cost of self-attention is also very high when processing high-resolution images. Unlike the method of generating attention matrices through matrix multiplication between query and key in self-attention (calculating the output of each pixel by summing up the weights of all other positions), convolutional modulation generates weights by using a depth convolution of $K \times K$ and re-weighting them through the Hadamard product.

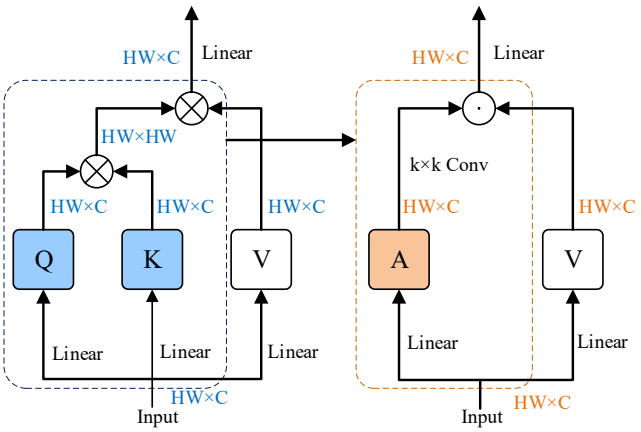


Figure 9. The operational process of self-attention mechanism and convolutional modulation.

The convolutional modulation method uses convolutional features extracted from deep convolution as weights to modulate the linear layers. Compared with residual blocks, ConvNeXt, and transformers, convolutional modulation uses convolution to establish relationships between network linear layers. This relationship is more memory efficient than self-attention, especially when processing high-resolution images. Compared with classical residual fast methods, convolutional modulation can adapt to input content determined by modulation operations.

For input tokens with an input length of N, the self-attention mechanism generates K (key), Q (query), and V (value) through a linear layer. The output is based on the weighted average of similarity score A, which is used to measure the relationship between each pair of input tokens as follows:

$$Attention(X) = A \cdot V, \tag{10}$$

$$A = Softmax(Q \cdot K^T). \tag{11}$$

Convolutional modulation uses a deep convolution kernel size of $K \times K$ for a given input token to perform convolutional feature modulation on the V value, simplifying self-attention and then calculating the output Z through the Hadamard product:

$$Z = A \odot V, \tag{12}$$

$$A = DConv_{k\times k}(W_1 X), \tag{13}$$

$$V = W_2 X. \tag{14}$$

Z is the output of convolutional modulation, and the output of each spatial position is the weighted sum of all pixels within a square area. $\odot$ is the Hadamard product, W_1 and W_2 are weight matrices generated by linear layers, and $DConv_{k\times k}$ represents deep convolutions with kernel size $K \times K$. By performing convolutional modulation operations, each spatial position (h, w) is correlated with all pixels within a $K \times K$ square region centered on (h, w), and information exchange between channels can be achieved through linear layers. The output of each spatial position is the weighted sum of all pixels within a square area.

2.4.2. MCT Module

The structure of the MCT module proposed in this paper is shown in Figure 10. YOLOv7 uses a 3 × 3 basic convolution module to extract features, but the receptive field is small and cannot capture more spatial contextual information. Using large kernel convolution is an effective method to help convolutional neural networks establish long-term relationships. By gradually increasing the depth of the convolution kernel, the model performance can be improved, but the number of parameters and computation will also increase accordingly. Considering that directly adding the Conv2Former module to YOLOv7 would result in a large computational load and make it difficult to meet the real-time requirements of dangerous goods detection, a lightweight module integrating Conv2Former, Conv2Former Block, is introduced to encode spatial feature information using convolutional modulation, solving the problem of the computational load of the self-attention mechanism increasing twice with the resolution of security inspection images. Using four Conv2Former Blocks with different kernel sizes to obtain important global location information improves the network's localization capability.

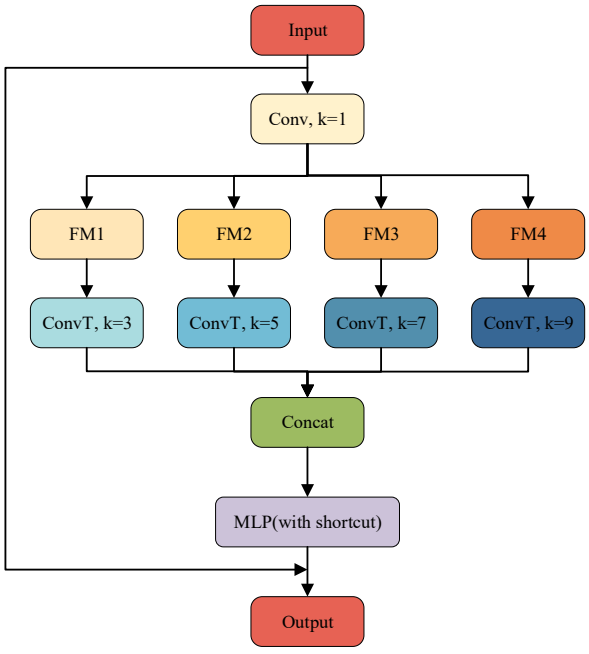


Figure 10. MCT module structure based on Conv2Former.

From Figure 10, it can be seen that the MCT module first adjusts the channel of the input image using 1 × 1 convolution, then uses the group convolution to segment the image into four sub-feature maps FM1, FM2, FM3, and FM4 (channel split), and then uses Conv2Former blocks with kernel sizes of 3 × 3, 5 × 5, 7 × 7, and 9 × 9 (k = 3, 5, 7, 9), respectively, to extract features from the sub-feature maps. After extracting features, the

sub-feature maps are concatenated and then combined with the input branch (to prevent gradient dispersion caused by network deepening) to undergo nonlinear transformation through MLP (multilayer perceptron) to better capture feature information at different positions before they are output.

By adding an MCT module, it is possible to accurately locate targets in high-density scenes in X-ray luggage images, alleviate the problem of small targets and overlap, prevent the loss of feature information at different scales, and improve detection accuracy.

3. Results

3.1. Dataset Description and Experimental Environment Settings

The X-ray image hazardous material detection experiment was conducted on three datasets, SIXray, CLCXray, and PIDray, as follows:

1. SIXray: The SIXray dataset [17] was released by the Pattern Recognition and Intelligent System Development Laboratory of the University of the Chinese Academy of Sciences and contains 1,059,231 X-ray screening images obtained from personal baggage scanning, of which 8929 X-ray screening images contain guns, knives, wrenches, pliers, or scissors—five kinds of contraband.
2. CLCXray: The CLCXray dataset [34] is a public dataset jointly released by the University of the Chinese Academy of Sciences, Tongji University, and Beijing University of Posts and Telecommunications. The dataset contains 9565 X-ray security images. There are 12 categories of contraband in the dataset, including five kinds of knives and seven kinds of liquid containers. The knives are specifically blades, daggers, knives, scissors, and Swiss army knives. The liquid containers are specifically cans, carton drinks, glass bottles, plastic bottles, vacuum bottles, spray cans, and tin cans.
3. PIDray: The PIDray dataset [35] is a large-scale X-ray security image dataset released in 2021, which was collected by Wang et al. using three different brands of security machines and contains 47,677 X-ray security images. There are a total of 12 categories of prohibited items in the dataset, namely batons, pliers, handcuffs, hammers, wrenches, lighters, scissors, knives, guns, powerbanks, sprayers, and bullets. The training set consists of 76,913 images, and the test set is divided into three small test sets classified as easy, hard, and hidden, corresponding to single-target, multiple-target, and intentionally hidden situations, respectively. The number of images is 24,758, 9746, and 13,069, respectively. This paper uses the easy test set to test the performance of the model.

Figure 11 shows X-ray images from the SIXray, CLCXray, and PIDray datasets, and Table 1 introduces the class distribution.

Table 1. Class distribution of SIXray, CLCXray, and PIDray datasets.

Label ID	Dataset					
	SIXray		CLCXray		PIDray	
	Name	Number	Name	Number	Name	Number
1	Gun	3131	Blade	3539	Baton	2399
2	Knife	1943	Dagger	988	Pliers	6814
3	Wrench	2199	Knife	700	Hammer	6229
4	Pliers	3961	Scissors	2496	Powerbank	8116
5	Scissors	983	Swiss Army Knife	1041	Scissors	7060
6			Cans	789	Wrench	6437
7			Carton Drinks	1926	Gun	3757
8			Glass Bottle	540	Bullet	2957
9			Plastic Bottle	5998	Sprayer	4227
10			Vacuum Cup	2166	HandCuffs	3388
11			Spray Cans	1077	Knife	5549
12			Tin	856	Lighter	6157

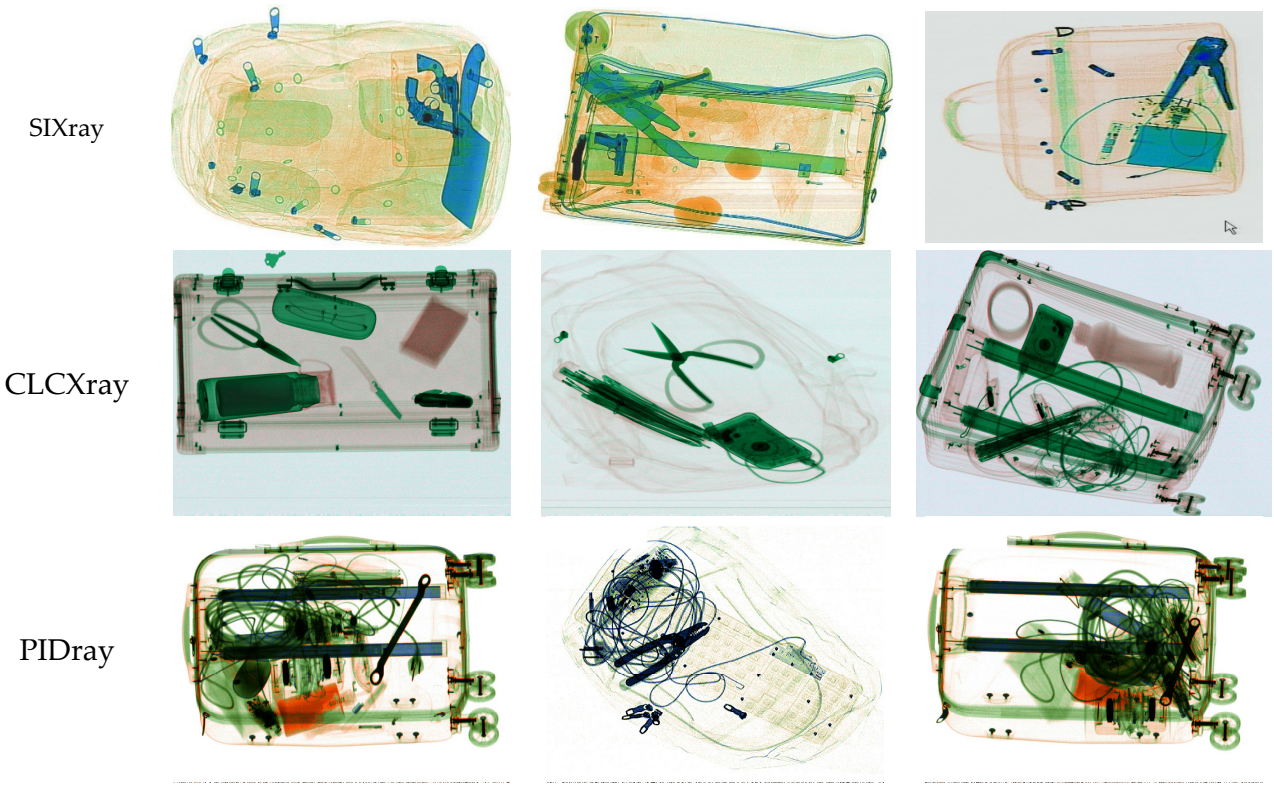


Figure 11. X-ray images of SIXray, CLCXray, and PIDray datasets.

The experimental environment is based on the Pycharm-professional-2019.3.2 compiler and Pytorch deep learning framework, using an Intel (R) Xeon (R) 4208 CPU @ 2.10 GHz processor for model training. The graphics card model is NVIDIA GeForce RTX 2080Ti. The maximum learning rate is 1×10^{-3}, the minimum learning rate is 1×10^{-6}, the epoch is set to 300, the iteration batch size value is 8, and MixUp and Mosaic are used as the main data augmentation methods. The Mosaic [36] probability is set to 1, and the MixUp [37] probability is set to 0.6. In order to further evaluate the performance of the model, this paper randomly divides X-ray security images containing prohibited items into training, validation, and testing sets in a ratio of 8:1:1.

3.2. The Evaluating Indicators

This paper used recall, F_1 score, precision, and mean average precision (mAP) as evaluation systems for object detection evaluation. The expressions for *precision*, *recall*, AP, mAP, and F_1 are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{AP} = \int_0^1 p(r)dr \tag{17}$$

$$\text{mAP} = \frac{\sum_{q=1}^{Q} \text{AP}(q)}{Q} \tag{18}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{19}$$

TP represents the number of correctly identified positive samples; *TN* is the correct number of negative samples for identification; *FP* represents the number of negative samples incorrectly identified as positive samples; and *FN* is the number of positive samples that are incorrectly identified as negative samples. The calculation of the *AP* value requires the P-R curve. mAP is the average of the values of all classes of *AP*. *Q* represents the total number of categories in object detection. The F_1 score can be calculated based on accuracy and recall.

3.3. Experimental Results and Analysis

In order to verify the detection performance of the improved YOLOv7 network on X-ray dangerous goods images, this paper compares the proposed improved algorithm with current mainstream object detection algorithms on the dataset SIXray: the Faster R-CNN (faster region-based convolutional neural network) algorithm, M2Det algorithm, SSD (single-shot detector) algorithm, YOLOv4, YOLOv5s, YOLOv7, and YOLOv8. A comparative experiment was conducted using mean average precision (mAP) as a measure of detection performance, and the results are shown in Table 2.

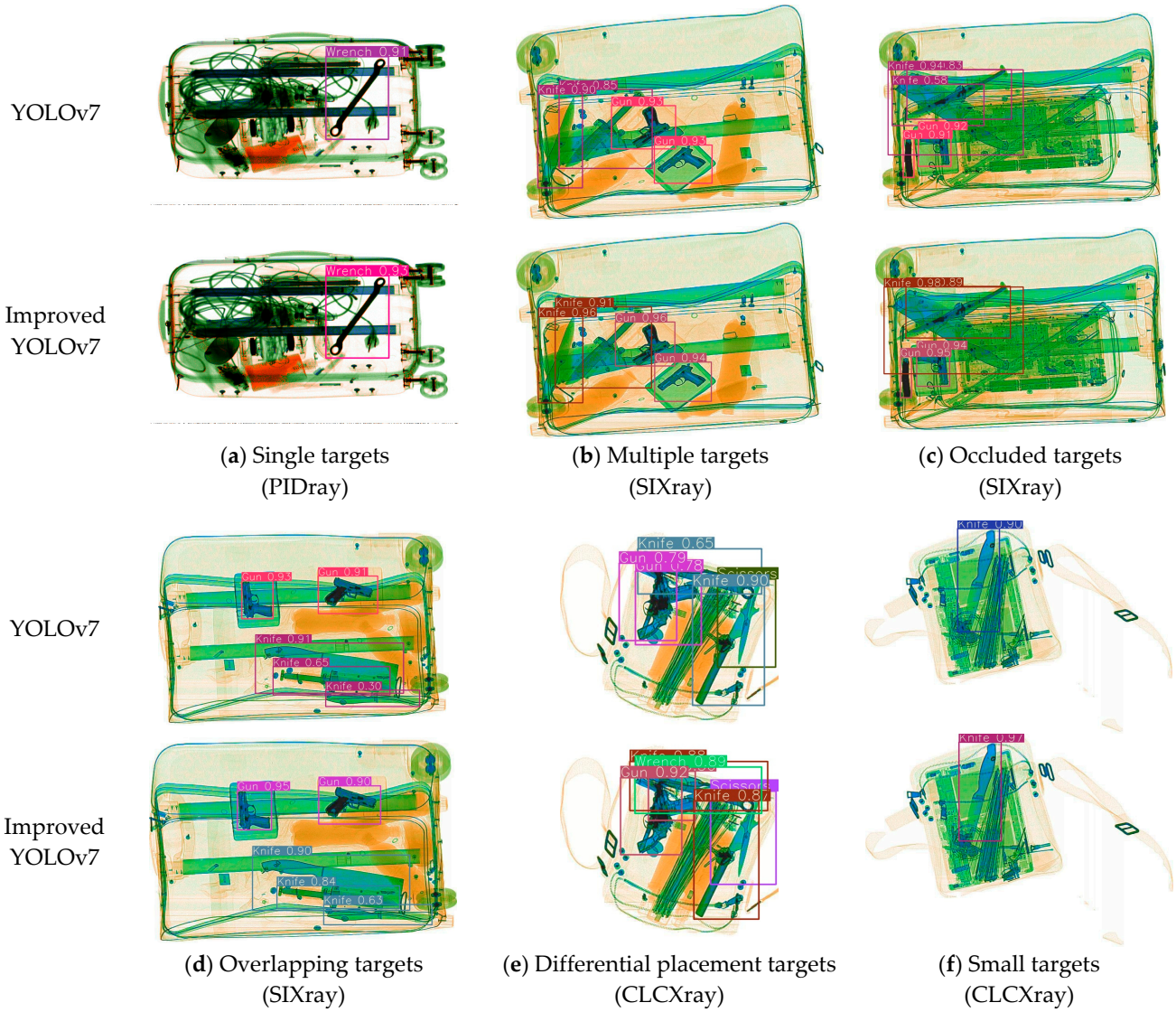
Table 2. Comparison of detection results of different models on the SIXray dataset.

Methods	AP (%)					mAP_0.5(%)
	Gun	Knife	Wrench	Pliers	Scissors	
Faster R-CNN	90.10	81.10	79.40	88.30	88.30	85.40
M2Det	95.49	75.70	70.17	83.00	82.96	81.47
SSD	94.91	77.87	74.82	84.51	82.69	82.96
YOLOv4	94.40	81.69	77.38	84.50	77.55	83.11
YOLOv5s	98.40	86.70	88.40	92.70	79.20	89.10
YOLOv7	98.60	90.00	92.40	93.80	83.30	91.60
YOLOv8n	98.47	88.90	93.74	94.22	84.50	91.96
Ours	100.00	92.50	94.60	97.30	96.80	96.30

mAP_0.5 represents the average accuracy of all categories when the IOU threshold is 0.5. IOU refers to the intersection-to-union ratio, which is used to measure the degree of overlap between predicted boxes and real boxes. The larger the ratio, the better the detection effect. From Table 2, it can be seen that the mAP of our model on the dataset SIXray is the highest at 96.3%, which is 10.9%, 14.83%, 13.34%, 13.19%, 7.2%, 4.7%, and 4.34% higher than Faster R-CNN, M2Det, SSD, YOLOv4, YOLOv5s, YOLOv7, and YOLOv8n, respectively. This model achieved the best AP in all five types of samples, fully demonstrating the effectiveness and advantages of this model in detecting dangerous goods in X-ray security images.

The detection performance of the improved model in the SIXray test set is shown in Figure 12, where (a) is a single sample image, (b) is a multiple sample image, (c) is an occluded sample image, (d) is an overlapping sample image, (e) is a differential placement sample image, (f) is a small sample image. It can be seen that our proposed method can accurately detect hazardous material targets from X-ray security images.

In order to further determine the detection situation of different types of hazardous materials, we analyzed the AP, accuracy precision, recall, and F1 scores of the improved method proposed in this paper for each hazardous material category in the SIXray dataset, as shown in Table 3. From Table 3, it can be seen that the AP, precision, recall, and F1 measures for the gun are the highest among all categories because the gun has the largest number of samples in the SIXray dataset. The accuracy, recall rate, and F1 score of the knife are relatively low, which is due to the flat and slender structure of the knife itself, which is prone to stacking with other items, resulting in missed detections.



(**a**) Single targets (PIDray) (**b**) Multiple targets (SIXray) (**c**) Occluded targets (SIXray)

(**d**) Overlapping targets (SIXray) (**e**) Differential placement targets (CLCXray) (**f**) Small targets (CLCXray)

Figure 12. Dangerous goods detection results by YOLOv7 and improved YOLOv7.

Table 3. Improved model performance analysis for each category.

Categories	AP (%)	Precision (%)	Recall (%)	F1 Measure
Gun	100.0	96.7	97.4	97.1
Knife	92.5	94.0	85.4	89.5
Wrench	94.6	91.8	89.3	90.5
Pliers	97.3	95.8	91.4	93.6
Scissors	96.8	94.4	88.0	91.1

On the dataset SIXray, the average value of various hazardous materials was used as the overall evaluation indicator to compare our method with the original YOLOv7, as shown in Figure 13. It can be seen that during the 300 rounds of training, the improved YOLOv7 algorithm performs well in precision, recall, mAP@0.5, and mAP@0.5:0.95. mAP_0.5 refers to calculating the average precision of all classes when the IOU threshold is 0.5. mAP_0.5:0.95 refers to calculating the average precision of all classes when the IOU threshold is set in steps of 0.05, ranging from 0.5 to 0.95. The four indicators show significant improvement compared to YOLOv7, indicating that our method is more effective in detecting dangerous goods in security checks and is more suitable for X-ray security check task scenarios.

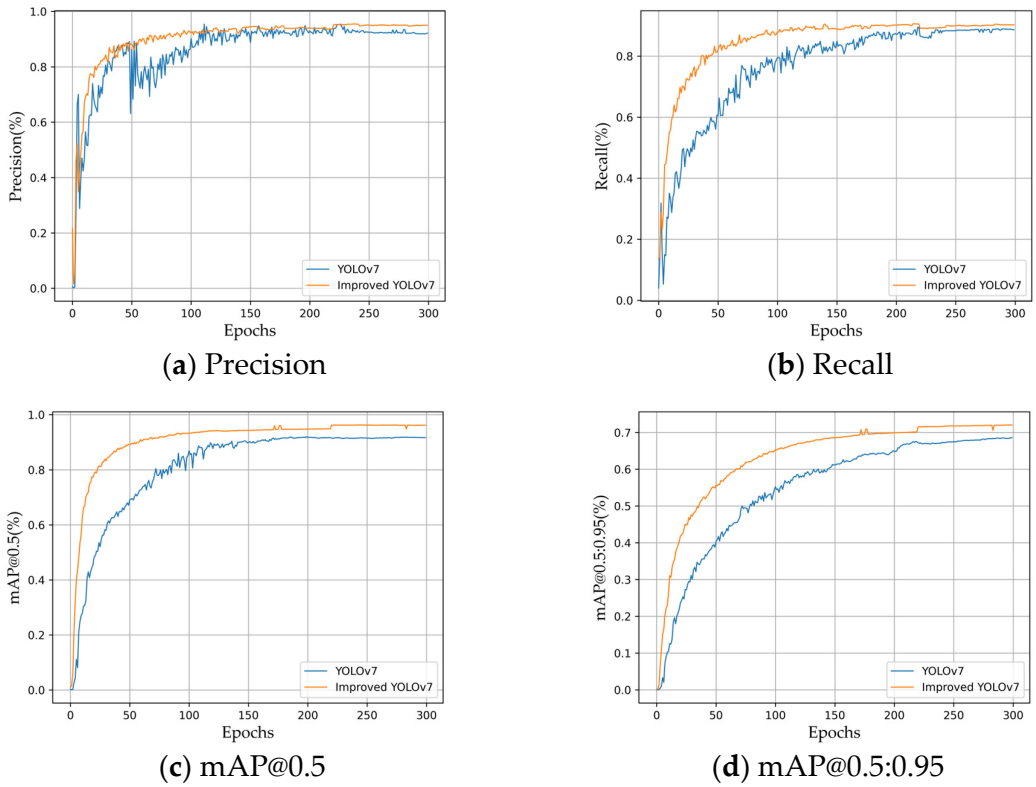


Figure 13. Performance comparison between YOLOv7 and the improved YOLOv7: (**a**) precision; (**b**) recall; (**c**) mAP@0.5; (**d**) mAP@0.5:0.95.

4. Discussion

In order to verify the robustness and universality of the improved model proposed in this article, which can be widely applied in X-ray security inspection tasks, it was validated on three X-ray security image datasets, SIXray, CLCXray, and PIDray, and compared with current mainstream object detection algorithms. The experimental results are shown in Table 4. From Table 4, it can be seen that the model proposed in this paper outperforms mainstream detection algorithms such as Faster R-CNN, SSD, YOLOv5s, and YOLOv8n on all three datasets. Compared to the baseline model YOLOv7, mAP achieved improvements of 4.7%, 2.7%, and 3.1%, respectively, proving the effectiveness of the improved method.

Table 4. Comparison experiments on three datasets.

Models	mAP@0.5[%]			FPS (SIXray)
	SIXray	CLCXray	PIDray	
Faster R-CNN	86.4	68.5	55.7	11
M2Det	84.2	66.7	58.6	44
SSD	83.9	64.6	57.8	38
YOLOv4	84.9	73.1	78.9	51
YOLOv5s	89.6	75.5	80.3	57
YOLOv7	91.6	76.6	81.6	69
YOLOv8n	92.0	77.2	82.5	74
Ours	96.3	79.3	84.7	65

In order to verify the effectiveness of the addition of various modules in the network proposed in this article, ablation experiments were conducted, and the experimental results are shown in Table 5. Method A represents the MPCN structure of the backbone network combined with coordinate attention and weighted bidirectional feature pyramid network, Method B represents the MPCF structure of the neck using dynamic snake convolution as a

cross-scale embedding layer, and Method C represents the multi-scale feature extraction module MCT combined with Conv2Former.

Table 5. Results of ablation experiment (SIXray dataset).

Models	AP[%]					mAP@0.5 (%)	mAP@0.5:0.95 (%)
	Gun	Knife	Wrench	Pliers	Scissors		
YOLOv7	98.6	90.0	92.4	93.8	83.3	91.6	68.6
YOLOv7 + A	98.3	90.5	94.0	93.9	85.5	92.5	69.4
YOLOv7 + AB	99.9	92.3	94.4	97.2	95.6	95.9	71.0
YOLOv7 + AC	100.0	92.3	94.8	96.9	97.2	96.2	71.7
YOLOv7 + ABC	100.0	92.5	94.6	97.3	96.8	96.3	72.1

The training process of the ablation experiment is shown in Figure 14. It can be seen from the figure that the improved methods proposed in this paper can effectively improve the accuracy of the model in detecting dangerous goods, proving the performance advantage of the proposed method.

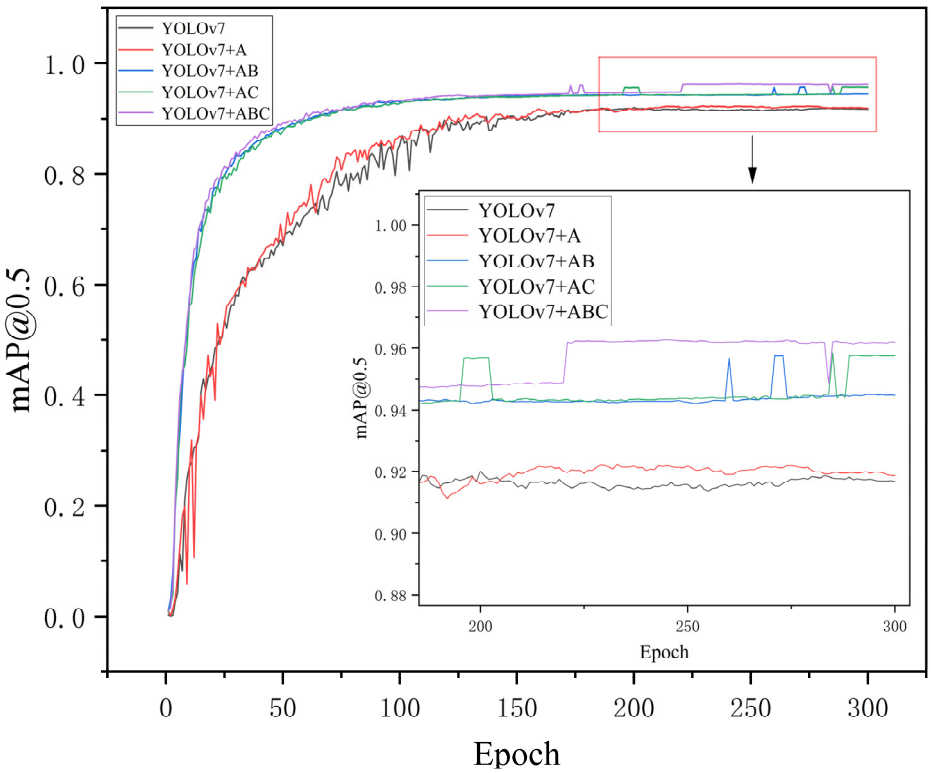
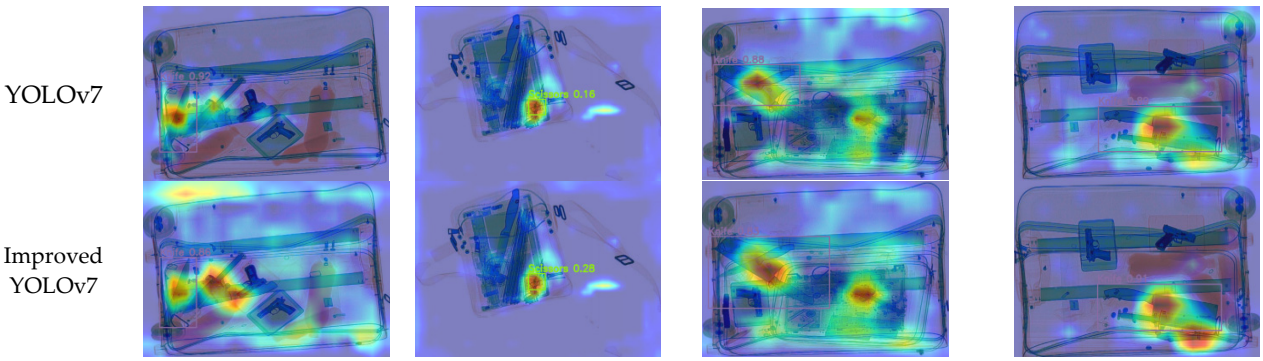


Figure 14. Visualization results of ablation experiment training process.

In order to further illustrate the role of the improved method in this article, based on the characteristics of X-ray security inspection images, typical case samples of four different situations were selected from the SIXray test set: ordinary target images, small target images, overlapping target images, and occluded target images. The GRAD-CAM algorithm [38] was used to generate comparative experimental heatmaps, as shown in Figure 15. It can be seen that for small target samples, the highlighted area of the heatmap can wrap around the entire area where the small target is located and the contour area where the target is located. The CA attention in the MPCN structure proposed in this paper can make the network's attention more focused on the small target area and more sensitive to the target's contour information. For occluded samples, the highlight range in the heatmap becomes wider, indicating that after adopting a dynamic serpentine convolutional

structure, the network can adaptively focus on the diverse feature information of hazardous material contour changes. For overlapping samples, the regions where hazardous materials are located at different scales are highlighted, indicating that with the help of Conv2Former, the network can obtain more critical features at different scales on a global scale. It can be seen that the method proposed in this article can effectively improve the network's detection performance.



(**a**) Ordinary targets (**b**) Small targets (**c**) Occluded targets (**d**) Overlapping targets

Figure 15. Heat map visual comparison of YOLOv7 and improved YOLOv7.

5. Conclusions

This paper takes the detection of hazardous materials in X-ray security images as the research object. Aiming at the current problems of complex image backgrounds, severe overlap of hazardous materials, and multi-scale pain points of targets, an improved X-ray security image hazardous material detection algorithm based on YOLOv7 is designed and implemented. The algorithm first combines a coordinate attention mechanism and weighted bidirectional feature pyramid network to improve the downsampling structure of the backbone, enhancing the feature extraction stage's ability to locate targets. Secondly, in the neck network, dynamic snake convolution is used as a cross-scale embedding layer to adaptively focus and learn the local feature information of hazardous materials, preventing the loss of some feature information caused by the decrease in feature map resolution. Finally, the designed MCT module uses Conv2Former's convolutional scaling method to simplify the self-attention mechanism, segment the feature maps of the input image into channels, and then extract features from each sub-map to overcome the problem of multi-scale differences in dangerous goods, enhancing the model's ability to extract global multi-scale features. In order to verify the universality of the method proposed in this article in X-ray security tasks, experiments were conducted on publicly available datasets, SIXray, CLCXray, and PIDray. Compared with YOLOv7, the mAP increased by 4.7%, 2.7%, and 3.1%, respectively. The experimental results show that our improved method can effectively improve the detection accuracy of the model and meet the requirements of X-ray security inspection tasks. However, the increase in computational and parameter requirements may cause difficulties in deploying to devices.

The aim of this paper is to improve the detection accuracy of hazardous materials in X-ray security inspection images. In the future, the improved method will be lightweight processed to further enhance the detection speed of the model and reduce the computational and parameter requirements, and a transportation hub security inspection and dangerous goods detection platform with a lightweight model as the core will be built to achieve intelligent security inspection.

Author Contributions: Conceptualization, Y.L., E.Z., X.Y. and A.W.; methodology, E.Z.; software E.Z.; validation Y.L. and E.Z.; writing—review and editing Y.L., E.Z., X.Y. and A.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Key Research and Development Plan Project of Heilongjiang (JD2023SJ19), the Natural Science Foundation of Heilongjiang Province (LH2023F034), the high-end foreign expert introduction program (G2022012010L), and the Key Research and Development Program Guidance Project of Heilongjiang (GZ20220123).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: SIXray: https://hyper.ai/datasets/18691, accessed on 15 February 2019; CLCXray: https://gitcode.com/GreysonPhoenix/CLCXray/tree/main/, accessed on 15 February 2022; PIDray: https://github.com/bywang2018/security-dataset, accessed on 15 August 2021.

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

References

1. Wang, Y.-L. Security Inspection Image Detection and Application Based on Deep Learning. *Mod. Inf. Technol.* **2021**, *5*, 82–85.
2. Han, B.-M.; Xi, Z.; Sun, Y.-J. *Summary of 2022 World Urban Rail Transit Operation Statistics and Analysis*; Urban Rapid Rail Transit: Beijing, China, 2022.
3. Bastan, M.; Yousefi, M.R.; Breuel, T.M. Visual words on baggage X-ray images. In Proceedings of the 2011 International Conference on Computer Analysis of Images and Patterns, Seville, Spain, 29–31 August 2011; pp. 360–368.
4. Turcsany, D.; Mouton, A.; Breckon, T.P. Improving feature-based object recognition for X-ray baggage security screening using primed visualwords. In Proceedings of the 2013 International Conference on Industrial Technology, Cape Town, South Africa, 25–28 February 2013; pp. 1140–1145.
5. Flitton, G.; Mouton, A.; Breckon, T.P. *Object Classification in 3D Baggage Security Computed Tomography Imagery Using Visual Codebooks*; Elsevier Science Inc.: Amsterdam, The Netherlands, 2015.
6. Mery, D.; Katsaggelos, A.K. A Logarithmic X-ray Imaging Model for Baggage Inspection: Simulation and Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 251–259.
7. Xing, X.-L.; Hu, H.; Xu, Y.-F. Security Inspection Image Interpretation and FPGA Implementation Method of Gray Scale Projection Algorithm. *Microcontroll. Embed. Syst.* **2018**, *18*, 42–44.
8. Russo, A.U.; Deb, K.; Tista, S.C.; Islam, A. Smoke Detection Method Based on LBP and SVM from Surveillance Camera. In Proceedings of the 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 8–9 February 2018; pp. 1–4.
9. Lyu, S.; Tu, X.; Lu, Y. X-ray image classification for parcel inspection in high-speed sorting line. In Proceedings of the 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 13–15 October 2018.
10. Akcay, S.; Breckon, T.P. An Evaluation of Region Based Object Detection Strategies 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. 1337–1341.
11. Mery, D.; Svec, E.; Arias, M.; Riffo, V.; Saavedra, J.M.; Banerjee, S. Modern Computer Vision Techniques for X-ray Testing in Baggage Inspection. *IEEE Access* **2017**, *47*, 682–692. [CrossRef]
12. Singh, B.; Li, H.; Sharma, A.; Davis, L.S. R-FCN-3000 at 30 fps: Decoupling Detection and Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 1081–1090.
13. Gu, L. Research and implementation of automatic cutlery recognition method based on X-ray security inspection image. In Proceedings of the 2020 3rd World Conference on Mechanical Engineering and Intelligent Manufacturing (WCMEIM), Shanghai, China, 4–6 December 2020; pp. 12–17.
14. Gaus, Y.F.A.; Bhowmik, N.; Akçay, S.; Guillén-Garcia, P.M.; Barker, J.W.; Breckon, T.P. Evaluation of a Dual Convolutional Neural Network Architecture for Object-wise Anomaly Detection in Cluttered X-ray Security Imagery. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
15. Li, D.; Hu, X.; Zhang, H.G.; Yang, J.F. A GAN based method for multiple prohibited items synthesis of X-ray security image. *Optoelectron. Lett.* **2021**, *17*, 112–117. [CrossRef]
16. Wang, Y.; Zhang, L. Dangerous goods detection based on multi-scale feature fusion in security images. *Laser Optoelect-Tronics Progress* **2021**, *58*, 152–159.
17. Miao, C.; Xie, L.; Wan, F.; Su, C.; Liu, H.; Jiao, J.; Ye, Q. SIXray: Alarge-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 (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: New York, NY, USA, 2019; pp. 2119–2128.
18. Tang, H.; Wang, Y.; Zhang, X. Dangerous goods detection algorithm by X-ray machine based on feature pyramid. *J. Xi'an Univ. Postand Telecommun.* **2020**, *25*, 58–63.
19. Zhang, Y.-K.; Su, Z.-G.; Zhang, H.-G. Multi scale detection of prohibited items in X-ray security inspection images. *J. Signal Process.* **2020**, *36*, 1096–1106.

20. Wei, Y.L.; Tao, R.S.; Wu, Z.J.; Ma, Y.; Zhang, L.; Liu, X. Occluded prohibited items detection: An X-ray security inspection benchmark and de-occlusion attention module. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 138–146.
21. Yang, F.; Jiang, R.; Yan, Y.; Xue, J.H.; Wang, B.; Wang, H. Dual-Mode Learning for Multi-Dataset X-ray Security Image Detection. *IEEE Trans. Inf. Forensics Secur.* **2024**. [CrossRef]
22. Lu, G.-Y.; Gu, Z.-H. Improved YOLOv3 security inspection algorithm for detecting dangerous goods in packages. *Comput. Appl. Softw.* **2021**, *38*, 197–204.
23. Wu, H.-B.; Wei, X.-Y.; Liu, M.-H. Combining dilated convolution and transfer learning to improve YOLOv4's X-ray security dangerous goods detection. *Chin. Opt.* **2021**, *14*, 1417–1425.
24. Dong, Y.-S.; Li, Z.-X.; Guo, J.-Y. An improved X-ray prohibited item detection model for YOLOv5. *Laser Optoelectron. Prog.* **2023**, *60*, 359–366.
25. Xianning, H.; Zhang, Y. ScanGuard-YOLO: Enhancing X-ray Prohibited Item Detection with Significant Performance Gains. *Sensors* **2023**, *24*, 102. [CrossRef] [PubMed]
26. Han, L.; Ma, C.; Liu, Y.; Jia, J.; Sun, J. SC-YOLOv8: A Security Check Model for the Inspection of Prohibited Items in X-ray Images. *Electronics* **2023**, *12*, 4208. [CrossRef]
27. Zhang, H.; Liu, B.-Y.; Gao, Y. X-ray security inspection recognition based on improved self attention neural network. *Laser J.* **2023**, *44*, 47–55.
28. Yang, P.; Yang, H.; Fang, C. Regional enhancement and multi feature fusion for identifying prohibited items in X-ray images. *J. Image Graph.* **2023**, *28*, 430–440.
29. Cheng, L.; Jing, C. X-ray image rotation target detection based on improved YOLOv7. *J. Graph.* **2023**, *44*, 324–334.
30. de Zarzà, I.; de Curtò, J.; Roig, G.; Calafate, C.T. LLM Multimodal Traffic Accident Forecasting. *Sensors* **2023**, *23*, 9225. [CrossRef]
31. Mukherjee, P.; Hou, B.; Lanfredi, R.B.; Summers, R.M. Feasibility of using the privacy-preserving large language model Vicuna for labeling radiology reports. *Radiology* **2023**, *309*, e231147. [CrossRef] [PubMed]
32. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. YOLO-World: Real-Time Open-Vocabulary Object Detection. *arXiv* **2024**, arXiv:2401.17270.
33. Minderer, M.; Gritsenko, A.; Houlsby, N. Scaling Open-Vocabulary Object Detection. *arXiv* **2023**, arXiv:2306.09683.
34. Zhao, C.; Zhu, L.; Dou, S.; Deng, W.; Wang, L. Detecting Overlapped Objects in X-Ray Security Imagery by a Label-Aware Mechanism. *IEEE Trans. Inform. Forensics Secur.* **2022**, *17*, 998–1009. [CrossRef]
35. Wang, B.; Zhang, L.; Wen, L.; Liu, X.; Wu, Y. Towars Real-World Prohibited Item Detection: A Large-Scale X-Ray Benchmark. In Proceedings of the IEEE/CVF Internatianal Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 5412–5421.
36. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
37. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
38. Zhu, B.-Y.; Liu, Z.; Zhang, J.-X. A COVID-19 detection algorithm combining Grad CAM and convolutional neural networks. *J. Front. Comput. Sci. Technol.* **2022**, *16*, 2108–2120.