X-ray Security Inspection Prohibited Items Detection Model based on Improved YOLOv7-tiny

Wenzhao Teng, Haigang Zhang, *a*nd Yujun Zhang

***Abstract*—X-ray security inspection for detecting prohibited items is widely used to maintain social order and ensure the safety of people's lives and property. Due to the large number of parameters and high computational complexity, most current object detection models are challenging to deploy on portable mobile security inspection devices. Therefore, this paper proposes an improved YOLOv7-tiny model for application in prohibited item detection tasks. Firstly, the feature extraction backbone network is replaced with the lightweight GhostNet network to reduce computational complexity and improve detection speed. Secondly, the FPN in the Neck is replaced with BiFPN, further reducing computational complexity and memory access through skip connections. Finally, a lightweight CA attention mechanism is embedded between the Backbone and Neck layers, and the Focal-EIoU Loss function is employed to enhance the detection capability for small-sized items. Experimental results on the SIXray public dataset show a 14.8% reduction in model parameters, a 19.7% reduction in computational complexity, and a 15.9% reduction in volume after the improvements. The detection speed increases from 82.4 to 90.2, and the detection accuracy for prohibited items reaches 90.3%. The experimental results demonstrate that the improved model achieves overall lightweighting while maintaining a high detection rate and improving detection speed.**

***Index Terms*—YOLOv7-tiny, GhostNet, BiFPN, CA attention mechanism, Focal-EIoU Loss.**

I. Introduction

SECURITY inspection is the first line of defense to safeguard people's lives and property. However, in real life, there are frequent extreme cases where criminals clandestinely carry prohibited items at airports and stations, threatening national security. Therefore, it is essential to strengthen security inspections of passengers and luggage in public places such as transportation hubs and crowded areas[1]. Traditional methods of detecting prohibited items primarily rely on X-ray transmission, mapping the pseudo-color images of detected items onto computer screens, and depending on trained security personnel for manual identification and inspection. However, during peak passenger flows and rapid pedestrian movements, security personnel may experience decreased attention due to fatigue, leading to instances of oversight and compromising people's safety and property[2]. Today, with the rapid development of computer vision technology, target detection models based on deep learning make intelligent analysis of X-ray security images possible.

With the continuous enrichment of deep learning theory and the continuous improvement of computer hardware performance, target detection technology is gradually being iterated and applied to security inspection equipment. However, most security inspection devices are edge terminal devices, and high-precision target detection models with large parameter and computational requirements are not conducive to deployment on these devices. Additionally, security inspection is a process that requires timeliness. For busy areas, a fast inspection process is crucial for maintaining order[3]. Therefore, lightweighting of detection models is particularly important while ensuring the accuracy of prohibited item detection. Given the relatively mature state of X-ray security inspection technology today, the specifications of corresponding security inspection machines tend to be standardized. Improving the hardware of security inspection machines is very difficult and costly. Therefore, without changing the hardware configuration, using lightweight target detection models is the most effective way to promote intelligent security inspection. Furthermore, in the process of detecting prohibited items in X-ray security images, the problem of missing small targets is common. The transmission effect of X-rays can cause small objects to be imaged unclearly. Traditional target detection models face challenges in detecting small prohibited items in pseudo-color images, making the results unsatisfactory[4]. In summary, seeking a lightweight and high-precision compatible target detection model for X-ray security image prohibited item detection is of great significance to assist security personnel in completing security inspection tasks.

Target detection models based on deep learning principles can be classified into two categories: two-stage and one-stage models. Two-stage models, predominantly represented by the R-CNN series, were pioneered by Girshick et al. in 2014 with the introduction of R-CNN[5]. This marked the first instance of the algorithmic concept of region proposal followed by classification and detection, significantly improving detection accuracy compared to traditional algorithms. In 2015, He et al. improved upon R-CNN by introducing Faster R-CNN, which drastically reduced the time spent on feature extraction by introducing region of interest pooling layers, further enhancing detection accuracy and efficiency[6]. Despite their high detection accuracy, two-stage models are characterized by a large number of parameters, resulting in lengthy algorithmic processes that are unsuitable for deployment on terminal devices. In contrast, one-stage models do not require region proposal selection; instead, they treat target detection as a regression task, enabling end-to-end

Manuscript received January 31, 2024; revised May 17, 2024. This work was supported by the Research Projects of Department of Education of Guangdong Province (2020ZDZX3082, 2023ZDZX1081, 2023KCXTD077, 2021ZDZX4064); School-level Project of Shenzhen Polytechinc University (6022310006K).

Wenzhao Teng a postgraduate student at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China (e-mail:tengwz619@163.com).

Haigang Zhang is a professor of Applied Artificial Intelligence of the Guangdong-Hong Kong-Macao Greater Bay Area, Shenzhen Polytechnic University, Shenzhen, China (e-mail:1834758165@qq.com).

Yujun Zhang is a professor at school of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China (e-mail:1997zyj@163.com).

detection. Prominent examples include the YOLO series, SSD[7], and RetinaNet[8] algorithms. One-stage models have fewer parameters, saving significant time during the detection process and making them suitable for deployment on terminal devices. The YOLO series algorithms have garnered widespread attention due to their excellent performance and effective balance between accuracy and speed. Therefore, this study will be based on the YOLO algorithm for research purposes.

As the demand for portable and high-precision detection in the security inspection industry continues to grow, significant progress has been made by experts and researchers in the field. Ren et al. proposed the LightRay model based on the YOLOv4 algorithm, using the lightweight MobileNetv3 network as the backbone for feature extraction[9]. They introduced a shallow feature enhancement network that integrates Feature Pyramid Network (FPN) and Convolutional Block Attention Module (CBAM), effectively addressing the detection of small-sized prohibited items in lightweight models. Cui et al. embedded the lightweight MobileNetViTv3 into the end of the YOLOv7 backbone network to capture comprehensive information, aiding in accurate positioning in high-density scenes, with the aim of balancing performance and practicality[10]. Sun et al. constructed the MobileNetv2 architecture for devices with limited computational capabilities, utilizing pointwise group convolutions and channel shuffling functions to reduce computational costs while maintaining accuracy[11]. The aforementioned research has significantly improved the performance of detection algorithms, laying a solid foundation for intelligent detection of prohibited items. However, there is still a need to further enhance the lightweight nature of models and the accuracy of prohibited item detection.

Therefore, the paper proposes an improved algorithm based on YOLOv7-tiny, effectively addressing the challenges of model lightweighting and the detection of small target prohibited items in contraband detection. In Chapter 5 of this paper, comparative experiments with other similar algorithms demonstrate the superiority of the improved algorithm. Additionally, the effectiveness of each improvement point is validated through ablation experiments.

II. Baseline model

In August 2022, Alexey Bochkovskiy and his team proposed the YOLOv7 series algorithm, which is the latest target detection network model and demonstrates significant advantages in terms of detection accuracy and speed compared to the previous YOLO series[12]. YOLOv7-tiny is characterized by its lightweight design, making it suitable for embedded devices and portable systems in resource-constrained environments. The network model of YOLOv7-tiny consists of three parts: the backbone, neck, and head. The backbone network is responsible for extracting features from input images, the neck fuses and processes the extracted features to obtain small, medium, and large-sized features. Finally, the fused features are passed to the detection head to produce the final output. YOLOv7-tiny incorporates the Efficient Lightweight Aggregation Network (ELAN) to enhance the feature extraction capabilities of the network while reducing computational complexity. In the backbone network's feature extraction, the MPConv module is introduced to expand the receptive field of the current feature layer and fuse it with information processed by conventional convolution, thereby improving the network's generalization. Additionally, the SPPCSPC module is applied at the end of the backbone network, introducing a series of convolution operations in parallel pooling to avoid issues such as image distortion. The neck network adopts the PANet pyramid structure for effective feature fusion between different layers[13]. Finally, in the prediction head, standard convolutions are used for channel adjustment to simplify the model structure while maintaining channel-adaptive effectiveness.

Compared to YOLOv7, YOLOv7-tiny sacrifices a certain degree of accuracy but demonstrates a noticeable advantage in terms of lightweight design and detection speed. In this paper, we aim to enhance the YOLOv7-tiny model and apply it to the detection of prohibited items in X-ray security inspection images. Due to the tight connection of each ELAN network with standard convolutions, there is computational redundancy in feature processing, leading to increased complexity in the network structure. Additionally, the model lacks the capability to extract features from small targets and items that are mutually occluded. Addressing these issues, the proposed model in this paper aims to further reduce model parameters and computational complexity while ensuring rich feature representation.

III. Improved model

This paper introduces an improved model based on YOLOv7-tiny, as shown in Figure 1, effectively addressing the challenges of model lightweighting and the detection of small target prohibited items. The improvements are outlined as follows:

(1) Replace the backbone network for feature extraction with the lightweight GhostNet network to reduce the number of parameters and computational complexity.

(2) Replace the FPN in the Neck network with BiFPN to reduce memory access and computational redundancy.

(3) Embed a lightweight Channel Attention mechanism between the Backbone and Neck to enhance the detection capability for small-sized targets.

(4) Replace the CIoU loss function with Local-EIoU Loss to improve the accuracy and robustness of the model detection.

A. GhostNet Network

Due to the use of regular convolutions for feature extraction in the backbone network of YOLOv7-tiny, the computational complexity increases, affecting the model's detection speed. Using the lightweight GhostNet network, channel interaction information is preserved through depthwise convolution operations, resulting in a significant reduction in floating-point computations compared to regular convolutions[14]. The network consists of multiple stacked BottleNeck structures, each of which is composed of Ghost Modules. The Ghost module is a core component of GhostNet, designed to reduce the computational cost of the network.

The GhostNet Bottleneck structure has two forms as shown in Figure 2. When the stride is set to 1, two Ghost Modules are directly used to increase the network depth.

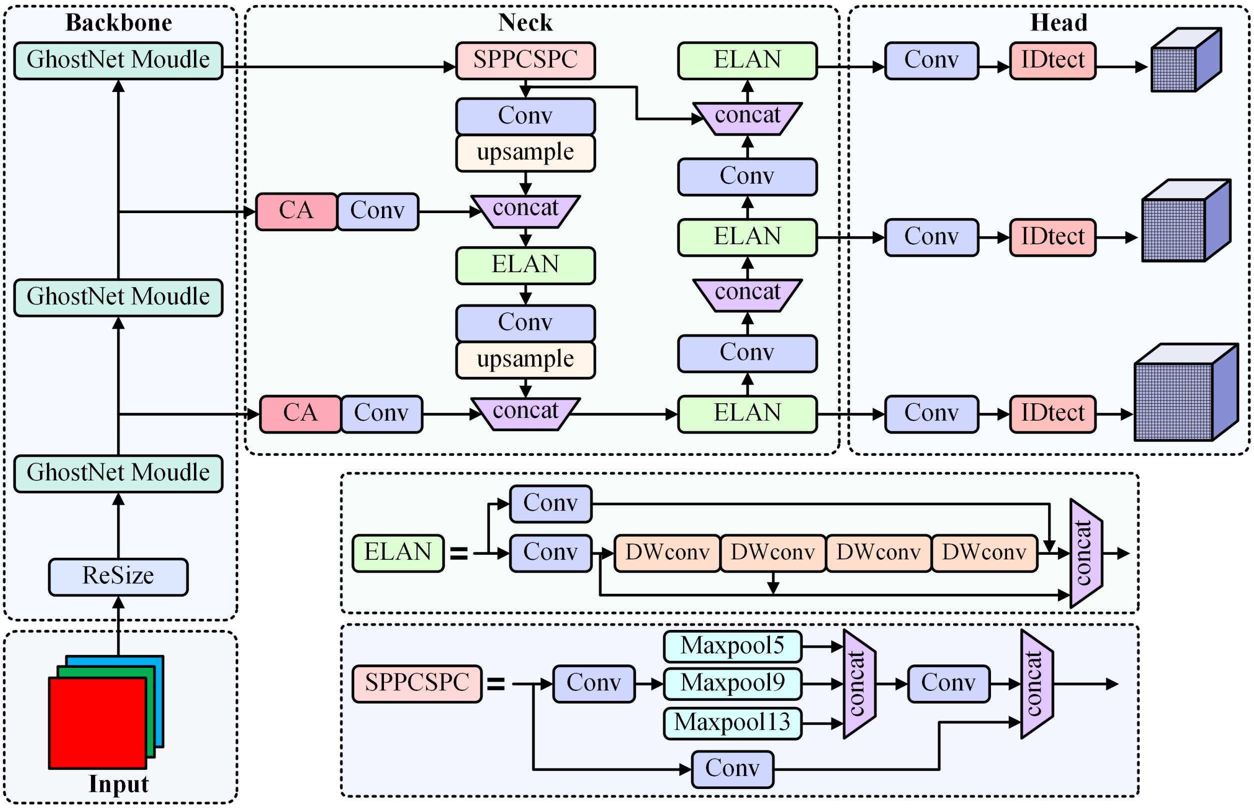

Fig. 1. The improved YOLOv7 network structure

When the stride is set to 2, a depth-wise separable convolution (DWConv) with a stride of 2 is added between two Ghost Modules to compress the width and height of the feature layer while meeting the requirements of the output channel number. In the residual connection, a DWConv with a stride of 2 and a regular convolution with a stride of 1 are added. Both forms combine convolutional and linear operations, where the linear operation helps remove redundant feature layers, reducing the computational cost associated with regular convolutions and thus improving the model's detection speed for contraband items.

In the Ghost Module, the input feature map undergoes a 1×1 Pointwise convolution operation to generate feature compression of the input feature layer. Then, DWConv is used to extract feature maps similar to the feature compression. Finally, these similar feature maps are integrated with the original feature compression to generate effective feature maps. In comparison to regular convolutions that use fixed-stride $3 \times 3 \times 3$ convolutional kernels to convolve $3 \times 3 \times 3$ input features, DWConv divides the input features into three layers. It uses three $3 \times 3 \times 1$ convolutional kernels corresponding to three different channels and extracts features from these layers through convolutional operations to obtain the output features of each layer. Simultaneously, it employs a $1 \times 1 \times 3$ convolutional kernel to extract feature information between each channel. Finally, the output features of each layer are fused with inter-channel features to obtain the final output features, effectively reducing the model's parameter count and computational complexity, thus improving both training and inference speeds. The principles of regular convolution and depth-wise separable convolution are illustrated in Figure 3.

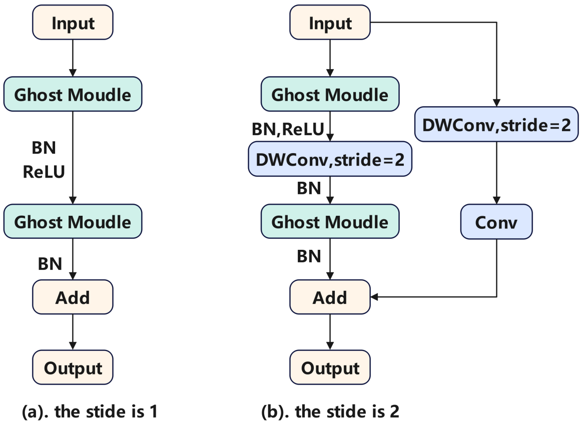

Fig. 2. The structure of GhostNet Moudle.

The later comparative ablation experiments have demonstrated that using GhostNet as the backbone network for feature extraction in YOLOv7-tiny reduces the model's parameter count and computational complexity, effectively achieving lightweighting of the security inspection model.

B. BiFPN

Real-time performance is a crucial requirement for object detection and poses a challenge when deploying algorithms on embedded platforms. To enhance the inference speed of the model and achieve real-time object detection, this paper replaces the traditional Feature Pyramid Network (FPN) with the Weighted Bidirectional Feature Pyramid Network

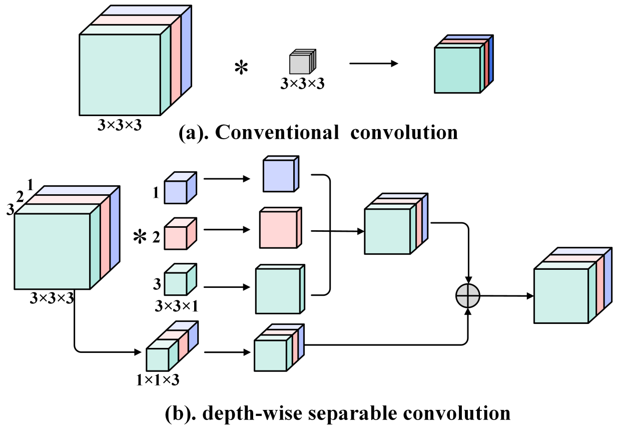
3×3×3
*
3×3×3
(a). Conventional convolution
1
2
3
3×3×3
1
*2
3
3×3×1
1×1×3
(b). depth-wise separable convolution

Fig. 3. Comparison of Conv and DWConv.

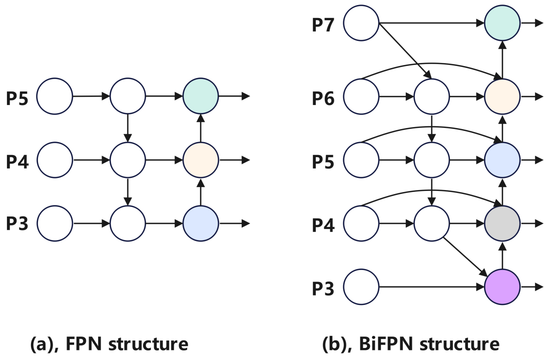

Fig. 4. Comparision of FPN and BiFPN.

(BiFPN)[15]. This adaptation aims to meet the real-time requirements of more efficient and compact embedded object detection.

BiFPN is a weighted bidirectional feature fusion structure that enables simple and fast multi-scale fusion. The structure is illustrated in Figure 4, and it is an improvement upon the traditional Feature Pyramid Network (FPN). On one hand, it removes the nodes located between the highest-dimensional feature layer and the lowest-dimensional feature layer. On the other hand, it adds residual edges connecting input and output feature maps for each feature layer located at the middle position.

Due to the varying resolutions of different input features, their contributions to the output features are uneven. To address this issue, BiFPN introduces additional weights for each input under different conditions. These weights are used to adjust the contribution to the output feature map. The weighted fusion is achieved through a fast normalized fusion, as shown in Formula 1. In terms of learning behavior and accuracy, this approach performs similarly to softmax-based optimizations, and it achieves a 30% speedup when running on a GPU.

$$O = \sum i \frac{\omega_i}{\varepsilon + \sum j\omega_j} \cdot I_i \tag{1}$$

In the formula, i and j represent the number of input feature maps at the node of feature fusion,I_i represents the input feature map matrix. To prevent the denominator from being zero, ε represents a constant. ω_i and ω_j represent the weights of each input feature map, where the initial range of weights is $0 < W_i < 1$, $0 < W_j < 1$.

With the stacking of layers, convolutional networks can capture richer semantic information. However, the reduction in feature map resolution results in the loss of positional information, which is particularly detrimental to the localization task in object detection. BiFPN achieves the comprehensive fusion of feature maps with different resolutions. It employs skip connections to achieve lightweighting, learning more critical feature information by introducing weights to the network. Considering the challenges in X-ray prohibited item detection, such as object occlusion and the detection of small-sized targets, precise identification of multiple prohibited items is essential. Therefore, by leveraging BiFPN to enhance the PANet in the improved YOLOv7-tiny model, the original feature information from the backbone network is directly introduced into the PANet of the neck network, effectively enhancing detection accuracy.

C. CA attention mechanism

Lightweight feature extraction backbone often lead to a decrease in model performance, especially for small object detection tasks. In order to make the model lightweight while ensuring the performance of prohibited item detection, we have introduced the CA module between the feature extraction backbone network and the feature fusion neck layer[16]. The CA attention mechanism not only captures inter channel information but also takes into account directionally relevant positional information, which can aid the model in better object localization and recognition. Moreover, the CA attention mechanism is sufficiently adaptable and lightweight to be seamlessly incorporated into the core structure of mobile networks. Experimental results reveal that the CA attention mechanism can effectively elevate model accuracy with only a minimal increase in computational cost. The structure of the CA attention mechanism is illustrated in Fig. 5.

In Fig. 5, *Residual* denotes the embeddable residual blocks, *X Avg Pool* signifies one-dimensional horizontal global pooling, and *Y Avg Pool* represents one-dimensional vertical global pooling. To avoid compressing all spatial information into channels and to capture more accurate positional information from distant interactions, global pooling is decomposed. This is achieved by using *X AvgPool* and *Y AvgPool* to decompose the input feature mapping into two one-dimensional feature coding processes. Dimensionally reduction is performed in both horizontal and vertical directions to obtain two feature vectors with dimensions $(H, 1)$ and $(1, W)$ respectively. These processes aggregate features along the horizontal and vertical directions respectively. The formulas for X-direction and Y-direction pooling are shown as follows.

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

$$Z_c^h(w) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{3}$$

Then, the extracted features are concatenated along the spatial dimension and mapped to the channel attention through a 1×1 convolution layer. The pooled result obtained

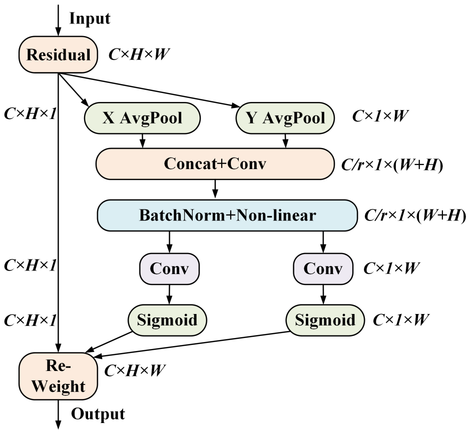

Fig. 5. CA structure diagram.

from Concat is subjected to shared 1×1 convolution, Batch Normalization layer, and non-linear activation function operations to generate an intermediate feature map. Subsequently, the intermediate feature map is separated along the spatial dimension, resulting in two tensors representing spatial information in the horizontal and vertical directions. Each of these tensors is then processed with the 1×1 convolution and combined with a sigmoid activation function to generate two attention weights. Finally, these two attention weights are multiplied and weighted with the original feature map, resulting in a feature map with coordinate awareness. It allows the network to focus more on relevant feature channels. The CA mechanism creates one branch carrying horizontal positional information and another branch carrying vertical positional information. This achieves the complete preservation of spatial positional information, enriching the position sensitivity of deep convolutions and enhancing the model's localization performance and detection accuracy. In improved model, the lightweight CA attention mechanism is embedded between the backbone network responsible for feature extraction and the neck network responsible for feature fusion. Subsequent ablation experiments confirm that adding the CA attention mechanism enhances the model's ability to detect prohibited items while maintaining its lightweight nature.

D. Focal-EIoU Loss

In the YOLOv7-tiny network, the loss function is composed of location loss, confidence loss and classification loss, as shown in Eqn. (4). In the proposed model, both confidence loss and classification loss are calculated using the BCEWithLogit loss function, while the location loss function is calculated using the Complete Intersection over Union (CIoU), as expressed in Eqn. (5).

$$Loss_{object} = Loss_{loc} + Loss_{conf} + Loss_{class} \tag{4}$$

$$Loss_{CIoU} = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{5}$$

where b and b^{gt} respectively denote the center coordinates of the predicted box and GT box. $\rho^2\left(b, b^{gt}\right)$ represents the Euclidean distance between the two center points, and c denotes the diagonal length of the minimum bounding rectangle for the two rectangles. Additionally, the loss is adjusted based on the aspect ratio αv. Due to CIoU only considering the overlapping area, center-point distance, and aspect ratio, without taking into account the bounding box, which is a crucial factor affecting the detection rate of prohibited items in security checks. Therefore, this paper introduces the Focal-EIoU loss function to replace the original loss function[17]. Focal-EIoU not only incorporates the advantages of CIoU but also focuses on high-quality bounding box detection, speeding up model convergence while improving the detection rate of prohibited items. The formula for calculating Focal-EIoU is shown in Eqn. (6).

$$L_{\text{EIoU}} = 1 - \text{IoU} + \frac{\rho^2\left(b, b^{gt}\right)}{(w_c)^2 + (h_c)^2} + \frac{\rho^2\left(w, w^{gt}\right)}{(w_c)^2} + \frac{\rho^2\left(h, h^{gt}\right)}{(h_c)^2} \tag{6}$$

where w and h represent the width and height of the minimum bounding box, respectively. The Focal-EIoU loss function, by increasing the similarity in aspect ratio, effectively reduces the genuine differences between $(w \times h)$ and $(w^{gt} \times h^{gt})$ through the regression form of focal loss. This is beneficial for improving the model's detection rate for small-sized prohibited items.

IV. Datasets and evaluation metrics

A. Experimental datasets

We apply SIXray dataset to verify the performance of the proposed model. It encompasses a total of 1,059,231 X-ray luggage images, out of which 8,929 annotated images are dedicated to object detection tasks[18]. These annotated images cover five categories of prohibited items: firearms, knives, wrenches, pliers, and scissors. These images were acquired through X-ray scanning of personal luggage at real security inspection locations. The employed dual-energy security inspection apparatus continues to be a prevalent technology for cargo luggage screening, extensively utilized in facilities such as airports, train stations, and subway stations. As a result, the SIXray dataset holds considerable research value and meets the demands of research inquiries. During the experiments, we partitioned these sample images into training, testing, and validation sets using a random ratio of 8:1:1. Examples of the five categories of prohibited items in SIXray are shown in Fig. 6.

B. Evaluation metrics

In this paper, the main evaluation metrics for the prohibited item detection in X-ray images include Precision (P), Recall (R), Average Precision (AP), Mean Average Precision (mAP), model Parameter count ($Params$), model computational complexity ($FLOPs$), Frames Per Second (FPS), and Model storage Size ($ModelSize$). As the evaluation metric for model accuracy, the mAP is divided into $mAP@0.5$ and $mAP@0.5:0.95$. $mAP@0.5$ represents the mAP value when the threshold is set to 50%. $mAP@0.95$ represents the mAP calculated as the threshold increases from 50% to 95% in increments of 5%, resulting in mAP values at different thresholds. In this study, we have chosen $mAP@0.5$ as the evaluation metric for model accuracy. A higher mAP value indicates higher overall model accuracy. The related metrics are calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{7}$$

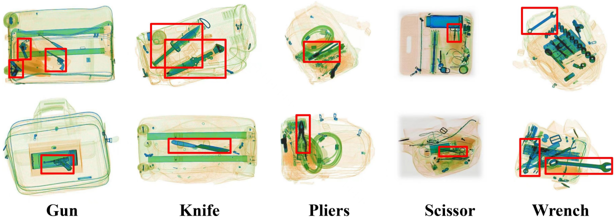

Fig. 6. Examples of prohibited items in the SIXray dataset.

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$AP = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$mAP = \frac{\sum_{n=1}^{Num(class)} AP(n)}{TP + TN + FP + FN} \tag{10}$$

where, TP represents the number of true positive samples correctly identified; TN represents the number of true negative samples correctly identified; FP represents the number of false positive samples incorrectly identified as positive; FN represents the number of false negative samples incorrectly identified as negative.

V. Experiment and Result Analysis

This paper designed two types of experiments to verify the performance of the proposed model. The first type is a comparative experiment, where the TinyRay model proposed in this paper is compared with typical object detection models. Additionally, to demonstrate the rationality behind selecting GhostNet as the feature extraction network, YOLOv7-tiny is used as the baseline model, and experimental comparisons are conducted using other lightweight network models as backbone networks. The second type is ablation experiments, where improvements are incrementally added to the baseline model to verify the effectiveness of each module in the proposed model.

A. Experimental configuration

The experiments were conducted on the Win10 operating system with the PyTorch 1.12 framework, and the GPU utilized was an NVIDIA RTX 3080. The batch size was set to 8, and the training was conducted for 300 epochs using Stochastic Gradient Descent (SGD) to adjust the network parameters. The image size was set to 640×640 pixels. The initial learning rate was set to 0.01, with a weight decay coefficient of 0.0005. The learning rate was adjusted using the cosine annealing algorithm.

B. Comparative experiment

To verify the effectiveness of the proposed model in this paper, a comparative experiment was conducted. In addition, the experiment selects current mainstream object detection algorithms, including Faster R-CNN[6], SSD[7] and DenseNet[19], as well as four lightweight variants from the YOLO series: YOLOv3-tiny[20], YOLOv4-tiny[21], YOLOv5s[22], YOLOv7-tiny, to make comparison with the proposed model. Faster R-CNN utilizes ResNet50 as the backbone network for feature extraction. The Region Proposal Network (RPN) is used to generate candidate boxes from features. After obtaining the feature matrix, predictions are made through fully connected layers. SSD employs VGG16 as the backbone network, with additional convolutional layers built on top of VGG16 to generate multiple feature maps for detection. DenseNet achieves feature reuse by densely connecting features across channels, resulting in a significant reduction in the number of parameters and computational costs. YOLOv3-tiny enhances the detection capability of objects of different sizes by combining Darknet53 architecture and reference space pyramid feature extraction module. YOLOv4-tiny took a significant leap forward by integrating the cross-stage local network CSP module into its foundational Darknet53 backbone. This strategic maneuver aimed to eliminate the challenge posed by the redundant duplication of gradient information during the intricate process of optimizing the network. YOLOv5s adopts the optimized CSPDarkNet53 as the backbone network and draws inspiration from the lightweight model EfficientDet. This addition transfers low-level localization features upward, allowing the pyramid to simultaneously possess semantic and localization information, thereby improving feature extraction. YOLOv7-tiny utilizes the Efficient-Aggregation Network and employs Spatial Pyramid Pooling in the neck network (SPPCSPC) to enhance feature extraction capabilities. The experimental results of the six algorithms and the improved algorithm proposed in this paper on the SIXray dataset are presented in Table 1.

From Table 1, it is evident that the YOLO series detection algorithms outperform the Faster R-CNN and SSD algorithms significantly in terms of lightweight design and detection speed for SIXray dataset. Among the YOLO algorithms, the novel YOLOv7-tiny algorithm not only achieves a mAP to 90.1%, which is the highest among all the mentioned algorithms, but also surpasses other algorithms in terms of lightweight design. The YOLOv7-tiny algorithm has 6.1 million parameters, a computation complexity of 13.2 billion operations, a model size of 12.3 megabytes, and a detection speed of 82.4 FPS. The performance of YOLOv7-tiny has laid a solid foundation for the model proposed in this paper. Through comparison, the proposed model reduces the model's parameter count, computation complexity, and model size by 14.8%, 19.7%, and 17.1%, respectively, resulting in 5.2 million parameters, 10.6 billion operations, and a model size of 10.2 megabytes. Simultaneously, the detection speed increases from 82.4 FPS to 90.2 FPS, while maintaining a detection accuracy of 90.3% for prohibited items. To sum up, proposed model effectively achieves a balance between lightweight design and detection accuracy.

Table 2 records the accuracy of detecting different prohibited items under various YOLO algorithms. With the updates in the YOLO series, the detection accuracy of the five prohibited item categories has improved to varying degrees. Notably, the proposed model in this paper achieves a detection accuracy as high as 99.0% for guns. At the same time, the detection accuracy of knife,pliers,scissors and wrench is as high as 88.3%, 90.5% ,85.5% and 88.0%, respectively, which is the highest value among many algorithms.Although the pliers and wrench have a certain degree of precision

TABLE I

RESULTS OF COMPARATIVE EXPERIMENTS ON DIFFERENT ALGORITHMS BASED ON SIXRAY DATASET

Algorithms	$Params\ (MB)$	$FLOPs\ (G)$	$Modelsize\ (MB)$	FPS (frames·s^{-1})	$mAP@0.5(\%)$
Faster R-CNN	135.7	125.3	511.3	17.1	87.2
SSD	41.1	387.0	270.2	47.1	83.4
DenseNet	27.2	124.6	112.0	40.5	77.4
YOLOv3-tiny	8.7	13.4	17.0	50.3	78.8
YOLOv4-tiny	6.0	13.7	23.8	58.7	81.5
YOLOv5s	7.2	15.9	14.8	67.5	88.2
YOLOv7-tiny	6.1	13.2	12.3	82.4	89.2
ours	**5.2**	**10.6**	**10.2**	**90.2**	**90.3**

TABLE II

DETECTION RESULTS AP FOR DIFFERENT CATEGORY OF PROHIBITED ITEMS

Algorithms	$AP(\%)$				
	gun	knife	pliers	scissors	wrench
YOLOv3-tiny	95.2	71.4	79.8	77.4	70.1
YOLOv4-tiny	96.0	74.5	82.6	80.2	74.5
YOLOv5s	97.7	83.7	88.4	86.3	84.8
YOLOv7-tiny	98.7	87.2	91.1	79.8	89.0
ours	99.0	88.3	90.5	85.5	88.0

TABLE III

COMPARISON OF DIFFERENT LIGHTWEIGHT BACKBONE NETWORKS ON YOLOV7-TINY.

Backbone	Params/ MB	FLOPs/ G	Model size/ MB	FPS/ frames·s^{-1}	mAP@0.5/ %
Mobilenetv2	6.4	18.0	15.4	80.3	85.9
Mobilenetv3	5.9	13.2	12.0	85.4	87.8
EffectiveVit	5.6	10.2	11.6	88.3	86.7
Shufflenet v1	6.5	14.2	14.6	79.8	88.1
Shufflenet v2	5.9	13.2	12.1	86.7	88.4
GhostNet	4.3	9.6	9.2	94.2	88.2

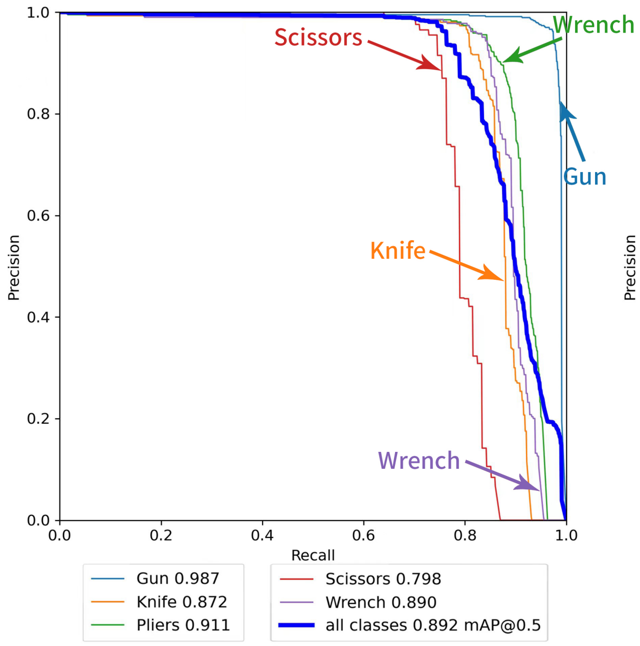

Fig. 7. Comparison of PR curves before and after improvement.

loss compared with the baseline model, the overall model proposed in this paper perfectly meets the task of contraband detection in security images.

To demonstrate the rationality behind selecting GhostNet as the feature extraction network, this paper conducted comparative experiments on the SIXray dataset by using other types of lightweight feature extraction backbones, namely Mobilenet v2, Mobilenet v3, EfficientVit, Shufflenet v1, and Shufflenet v2, based on the YOLOv7-tiny model. MobileNet v2 improves the network's representational capacity on a lightweight basis through the use of linear bottlenecks and inverted residuals. Mobilenetv3, proposed by A. Howard et

TABLE IV
RESULTS OF ABLATION EXPERIMENTS ON THE SIXRAY DATASET

Model	Params/ MB	FLOPs/ G	Model size/ MB	FPS/ frames·s^{-1}	mAP@0.5/ %
YOLOv7-tiny	6.1	13.2	12.3	82.4	89.2
YOLOv7-tiny+A	4.3	9.6	9.2	94.2	88.2
YOLOv7-tiny+A+B	5.2	10.2	10.0	91.3	89.0
YOLOv7-tiny+A+B+C	5.2	10.4	10.2	90.6	89.9
YOLOv7-tiny+A+B+C+D	5.2	10.6	10.2	90.2	90.3

al. improves upon the Mobilenetv2 structure with an inverted residual design containing DWConv and linear bottleneck. EfficientVit is based on the EfficientViT block, each block is composed of a sandwich structure and a cascaded group attention mechanism. The authors achieve a more efficient balance of channel, block, and stage quantities through parameter redistribution. Shufflenet v1, proposed by Zhang et al. introduces Pointwise group convolution and channel shuffle to enable the network to have more channels. This helps extract more information during the encoding stage while reducing computational complexity. Shufflenet v2 further optimizes the network by introducing channel split and feature reuse, reducing the number of parameters while enhancing detection accuracy. Table 3 shows the comparative experimental results of using GhostNet compared to the aforementioned lightweight networks on the basis of YOLOv7-tiny.

The comparative experiments in Table 3 reveal that using Mobilenetv3 and Shufflenet v2 as the feature extraction backbone network for YOLOv7-tiny results in minimal fluctuations in parameters, computational complexity, and detection speed, with less-than-ideal performance. Shufflenet v1's various evaluation metrics are even worse than the original model, failing to guarantee satisfactory detection results. Although the EfficientVit network has lower parameter count and computational complexity, it incurs significant accuracy loss. In contrast, using GhostNet as the backbone network for feature extraction, the number of parameters decreased from 6.1M to 4.3M, the computational complexity decreased from 13.2G to 9.6G, the model volume decreased from 12.3M to 9.2M, the FPS increased from 82.4 to 94.2, and the detection accuracy remained at 88.2%. Based on the original YOLOv7-tiny, it can effectively reduce the number of parameters, computational complexity and model size, and improve the detection speed without losing too much accuracy. Therefore, adopting GhostNet as feature extraction network and improving it is a more reasonable choice for this paper.

C. *Ablation experiments*

To validate the effectiveness of each improvement component on network performance, the original YOLOv7-tiny is taken as the baseline model, and ablation experiments are conducted by gradually incorporating each improvement module into YOLOv7-tiny. The experimental results are shown in Table 3, where Method A represents replacing the backbone network with the lightweight GhostNet network, method B means to replace FPN structure in Neck network with BiFPN structure, method C means to embed lightweight CA attention mechanism between Backbone layer and Neck layer, and method D means to replace CIoU loss function with Local-EIoU loss function It is obvious that "YOLOv7-tiny+A+B+C+D" in Table IV represents the proposed model in this paper.

The ablation experiments reveal the following observations. In Experiment A, after replacing the backbone network with the lightweight GhostNet network, there is a significant decrease in parameters, computational complexity, and model size. The parameter count and computational complexity decrease by 29.6% and 27.3% respectively, with a 25.3% reduction in model size. The detection speed increases from 82.4 FPS to 94.2 FPS, while a 1.9% decrease in detection accuracy is observed. In experiment B, the FPN structure in Neck layer is replaced by BiFPN structure. Although the number of parameters and computational complexity are increased, the precision is effectively improved by 1.7%. In experiment C and experiment D, the lightweight CA attention mechanism and modified loss function were embedded, and the detection accuracy was improved by 1.2% and 0.4% respectively without affecting the lightweight of the model. In conclusion, the ablation experiments demonstrate that the improved model achieves model lightweighting and enhances the prohibited item detection speed while maintaining a high detection accuracy.

D. *Visual analysis*

To visually illustrate the differences between the model before and after improvement, four images were selected to generate detection result visualizations using a visualization algorithm, as shown in Figure 8. A comparison reveals that the improved model in this study is able to detect prohibited items that were not detected by YOLOv5s, and there is an overall improvement in the detection accuracy for each category of prohibited items. In summary, the improved model in this study maintains excellent detection performance while being lightweight.

VI. CONCLUSION

This paper proposes a lightweight object detection model based on YOLOv7-tiny framework. Drawing inspiration from the lightweight network GhostNet, the backbone network is improved, which can effectively reduce model parameters and computational complexity while accelerating prohibited item detection speed. BiFPN structure is adopted in the neck network, and the detection capability of contraband is enhanced by embedding lightweight CA attention mechanism and Local-EIoU loss function, while maintaining lightweight design.Experimental results demonstrate that the improved model successfully achieves a balance between lightweight design and detection accuracy for prohibited item detection tasks in X-ray images.

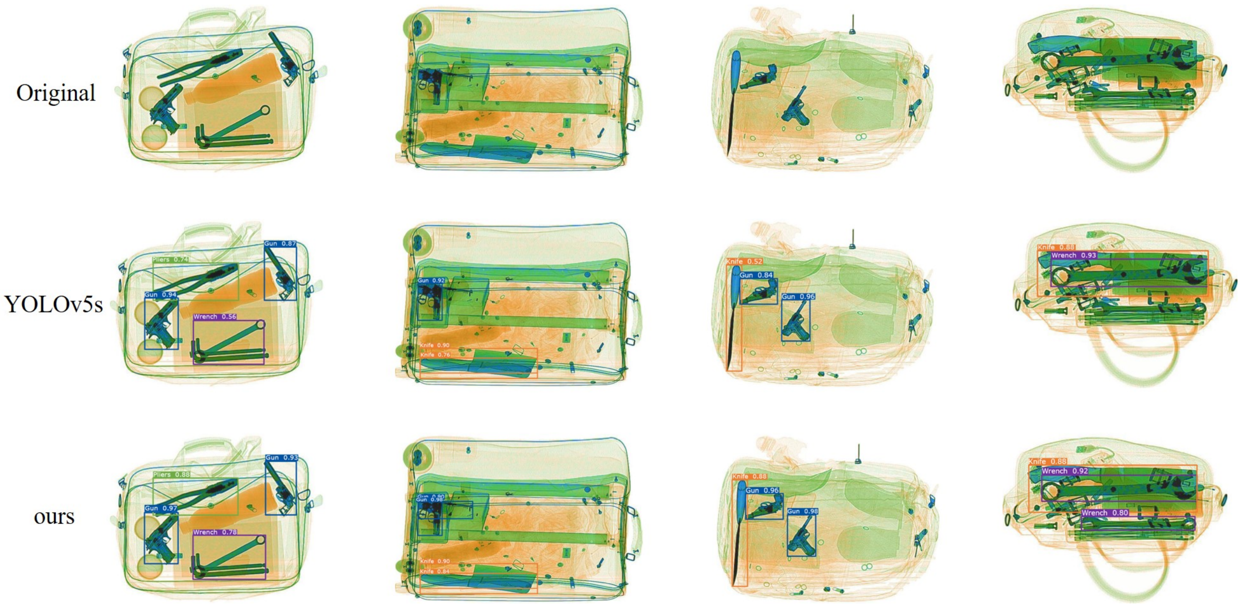

Fig. 8. Visualization of case.

References

[1] Q. Wang and T. P. Breckon, “Contraband materials detection within volumetric 3d computed tomography baggage security screening imagery,” in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, Dec. 2021.

[2] S. Akcay and T. Breckon, “Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging,” *Pattern Recognition*, vol. 122, p. 108245, Feb. 2022.

[3] G. Batsis, I. Mademlis, and G. T. Papadopoulos, “Illicit item detection in x-ray images for security applications,” in *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, Jul. 2023.

[4] L. Shen, W. Cui, Y. Tao, T. Shi, and J. Liao, “Surface defect detection algorithm of hot-rolled strip based on improved yolov7.” *IAENG International Journal of Computer Science*, vol. 51, no. 4, 2024.

[5] R. Girshick, “Fast r-cnn,” in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Dec. 2015.

[6] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, p. 1137–1149, Jun. 2017.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[8] M. Cheng, J. Bai, L. Li, Q. Chen, X. Zhou, H. Zhang, and P. Zhang, “Tiny-retinanet: a one-stage detector for real-time object detection,” in *Eleventh International Conference on Graphics and Image Processing (ICGIP 2019)*, Z. Pan and X. Wang, Eds. SPIE, Jan. 2020.

[9] Y. Ren, H. Zhang, H. Sun, G. Ma, J. Ren, and J. Yang, “Lightray: Lightweight network for prohibited items detection in x-ray images during security inspection,” *Computers and Electrical Engineering*, vol. 103, p. 108283, 2022.

[10] C. Liqun and J. Yaqin, “Improved x-ray prohibited items detection algorithm for yolov7,” in *2023 IEEE 6th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*. IEEE, Dec. 2023.

[11] P. Sun, H. Zhang, J. Yang, and D. Wei, “Mobilevit based lightweight model for prohibited item detection in x-ray images,” in *Asian Conference on Pattern Recognition*. Springer, 2023, pp. 45–58.

[12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2023.

[13] K. Liu, Q. Sun, D. Sun, L. Peng, M. Yang, and N. Wang, “Underwater target detection based on improved yolov7,” *Journal of Marine Science and Engineering*, vol. 11, no. 3, p. 677, Mar. 2023.

[14] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, “Ghostnet: More features from cheap operations,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2020.

[15] J. Chen, H. Mai, L. Luo, X. Chen, and K. Wu, “Effective feature fusion network in bifpn for small object detection,” in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, Sep. 2021.

[16] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2021.

[17] Q. Weimin, C. Hangong, Y. Yuting, and Y. Guoshuai, “Indoor object recognition based on yolov5 with eiou loss function,” in *Third International Conference on Advanced Algorithms and Signal Image Processing (AASIP 2023)*, K. Subramaniam and P. Loskot, Eds. SPIE, Oct. 2023.

[18] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. Ye, “Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2119–2128.

[19] G. Huang, S. Liu, L. v. d. Maaten, and K. Q. Weinberger, “Condensenet: An efficient densenet using learned group convolutions,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2018.

[20] H. Gong, H. Li, K. Xu, and Y. Zhang, “Object detection based on improved yolov3-tiny,” in *2019 Chinese Automation Congress (CAC)*. IEEE, Nov. 2019.

[21] S. Ali, A. Siddique, H. F. Ates, and B. K. Gunturk, “Improved yolov4 for aerial object detection,” in *2021 29th Signal Processing and Communications Applications Conference (SIU)*. IEEE, Jun. 2021.

[22] T.-H. Wu, T.-W. Wang, and Y.-Q. Liu, “Real-time vehicle and distance detection based on improved yolo v5 network,” in *2021 3rd World Symposium on Artificial Intelligence (WSAI)*. IEEE, Jun. 2021.