STRay: A Model for Prohibited Item Detection in Security Check Images

Wenzhao Teng, Yujun Zhang

Abstract—**Addressing issues such as mutual occlusion of items and small scale of prohibited items in X-ray security inspection image detection, we propose an improved X-ray contraband detection model based on YOLOv7 named STRay. Firstly, in the backbone network, the model employs Swin Transformer, applying a sliding window multi-head self-attention mechanism to suppress background interference, enabling the network to focus more on contraband items and reducing the false negative rate. Secondly, conventional convolutions in E-ELAN are replaced with deformable dilated convolutions, adjusting the convolutional kernel's shape by learning sampling offsets to better match the contours of contraband items and effectively address mutual occlusion issues. Lastly, the detection head in the head section is replaced with an Efficient decoupled detection head, decoupling separate feature channels for localization and classification tasks, thereby enhancing the classification and localization capabilities for small-scale contraband items. The proposed model is tested on large datasets SIXray, OPIXray, and PIDray, achieving mAPs of 95.3%, 88.8%, and 83.1% respectively, effectively improving contraband detection capabilities while maintaining fast detection speeds. Compared to current mainstream models, it demonstrates certain advancements, providing excellent technical support for ensuring public safety.**

Index Terms—**security inspection, YOLOv7, Swin Transformer, deformable dilated convolution , efficient decoupled detection head.**

I. Introduction

In recent years, with the prosperity of the economy and advancements in science and technology, there has been high-quality development of public infrastructure globally. Particularly, the construction of comprehensive and three-dimensional public transportation systems such as aviation and high-speed rail has effectively met the personalized and diversified travel needs of the people, enhancing the convenience of travel. However, behind the enjoyment of these conveniences lies significant security risks. Extreme cases of individuals carrying prohibited items illegally at airports and on high-speed trains threaten national and social security and are not uncommon. Therefore, it is crucial to strengthen security checks on passengers and their luggage in transportation hubs and crowded public places. Currently, luggage security checks primarily rely on security personnel to intelligently judge and identify pseudo-color images generated by X-ray inspection machines. However, during peak passenger flows and rapid transit, security personnel may experience decreased attention due to fatigue, leading to missed inspections and compromising the safety of people's lives and property[1]. In summary, seeking a high-precision X-ray security image prohibited item detection model to assist security personnel in completing security checks is of significant research importance.

Based on deep learning, object detection models can be classified into two categories according to their algorithm principles: two-stage and one-stage models. Two-stage models, represented by the R-CNN series, were initiated in 2014 when Girshick et al. proposed R-CNN, which introduced the idea of region proposal followed by classification and detection, significantly improving detection accuracy compared to traditional algorithms[2]. The following year, He et al. improved R-CNN with Faster R-CNN by introducing region of interest pooling layers, greatly reducing the time required for feature extraction[3]. However, two-stage models have a large number of parameters, leading to a time-consuming algorithm process that is not suitable for deployment on terminal devices. In contrast, one-stage models, such as the YOLO series, SSD[4], and RetinaNet[5], do not require region proposal and treat object detection as a regression task, achieving end-to-end detection. One-stage models have fewer parameters and save a significant amount of time during the detection process. Among them, the YOLO series has garnered widespread attention due to its outstanding performance and effective balance between accuracy and speed. With continuous updates, YOLO models have surpassed two-stage models in accuracy and are suitable for deployment on terminal devices. Therefore, this paper will conduct research based on the YOLO model.

The main task of prohibited item detection in X-ray security images is to identify the types of prohibited items and locate their positions in pseudo-color images. In actual scenarios, security checks encounter many inevitable challenges. Firstly, in X-ray security images, the cluttered arrangement of items is a common problem. The transmission of X-rays causes items to overlap, forming occlusions between them. Additionally, different materials create a complex background with multiple overlapping colors in pseudo-color images[6]. Traditional object detection models struggle to handle the cluttered and overlapping items in pseudo-color images, resulting in unsatisfactory performance. Secondly, luggage and backpacks contain various items of different sizes and types. For prohibited item detection, this presents a multi-scale, multi-target detection problem, where differences in scale can lead to the model overlooking small-scale targets, resulting in missed detections.

Facing the aforementioned challenges, deep learning-based object detection models have provided a solid theoretical foundation for the development of intelligent security

Manuscript received March 30, 2024; revised August 6, 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 is 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).

Yujun Zhang is a professor at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China (e-mail:1834758165@qq.com).

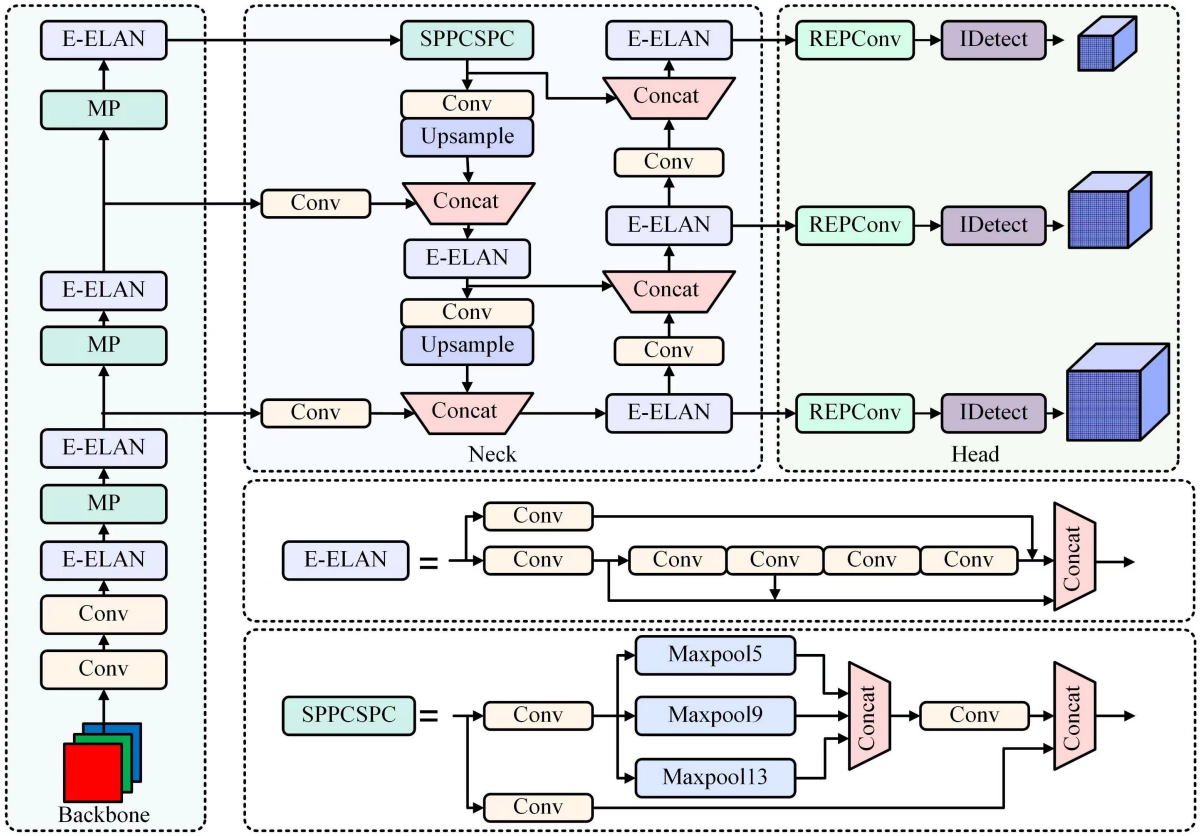

Fig. 1. YOLOv7 Network Structure

checks, and experts and researchers in the security industry have achieved significant results. In response to the difficulty of deploying security models on portable inspection devices, Ren et al. proposed the LightRay model based on the YOLOv4 algorithm, utilizing MobileNetv3 as the feature extraction backbone network[7]. They introduced a shallow feature enhancement network that combines Feature Pyramid Network (LFPN) and Convolutional Block Attention Mechanism (CBAM) to strengthen the feature extraction of small target objects while lightweighting the model. Addressing the issue of prohibited items occluding each other, Shao et al. applied a foreground-background separation model to adaptively learn the foreground features containing prohibited items[8]. They separated the foreground objects and background in sample images, enabling the network to focus more on learning foreground objects and suppressing background interference. Wei proposed the De-occlusion Attention Module (DOAM), emphasizing the extraction of edge and material information of prohibited items from an attention perspective to enhance the detection capability of prohibited items in X-ray images[9]. The aforementioned research has improved the performance of detection algorithms to varying degrees, laying a solid foundation for the intelligent detection of prohibited items. However, the detection of prohibited items in real-world scenarios still faces challenges such as occlusion between prohibited items and weak targets, necessitating further improvement in accuracy and robustness.

Therefore, this paper designs an improved model based on YOLOv7 named STRay. The proposed model is validated and tested on three large public datasets, effectively addressing the issues of occlusion between items and the omission of small-sized targets in contraband detection. In Chapter Five of this paper, through comparative experiments with other similar algorithms, the superiority of the improved algorithm is demonstrated, and the effectiveness of each improvement point is verified through ablation experiments.

II. Propaedeutics

The YOLOv7 series algorithm, proposed by Alexey Bochkovskiy et al., presents notable improvements over previous iterations of the YOLO series in terms of both detection accuracy and speed[10]. Among real-time object detectors capable of achieving over 30 frames per second on the GPU V100, YOLOv7 demonstrates superior accuracy and detection rate. YOLOv7's architecture consists of three main components: the backbone network, the neck, and the head. The backbone extracts features from input images, while the neck merges these features into small, medium, and large-sized representations. These combined features are then forwarded to the detection head, which produces the final detection results. Figure 1 provides a visual depiction of YOLOv7's architecture.

The core structure of YOLOv7's backbone network primarily comprises convolutional layers, Expandable Efficient Linear Aggregation Network (E-ELAN) modules, MPConv modules, and SPPCSPC modules. The E-ELAN module, which is an extension of the original ELAN, leverages techniques such as expansion and shuffling to enhance the network's learning capacity without disrupting the gradient path. This results in an increased feature extraction capability for the network. MPConv enhances the feature layer's receptive

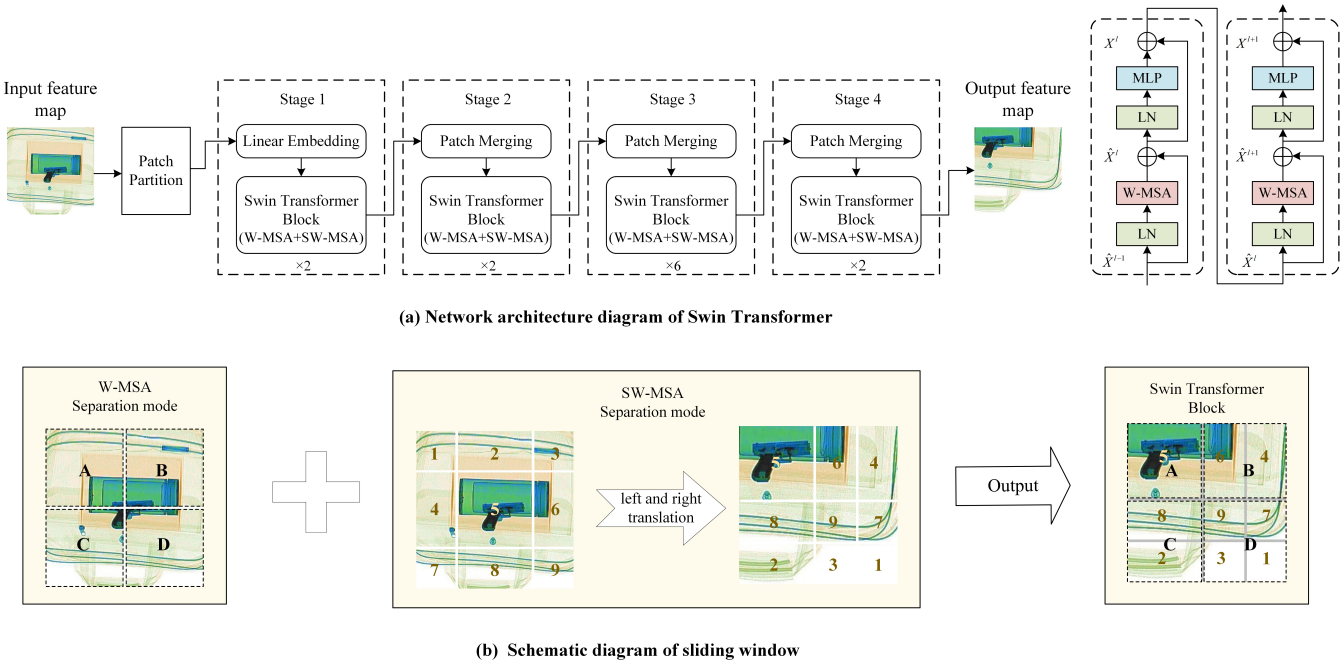

Fig. 2. Network architecture and principle of Swin Transformer attention mechanism.

field and integrates it with traditional convolution outputs, boosting the network's ability to generalize. Towards the end of the backbone network, the SPPCSPC module applies multiple convolution operations alongside parallel pooling to combat image distortion and extract non-redundant features in CNNs. In YOLOv7's feature fusion Neck network, akin to prior YOLO versions, it employs the Path Aggregation Feature Pyramid Network (PAFPN) structure [11] and incorporates the E-ELAN module, effectively aggregating information across various network paths or feature pyramids with adaptive receptive fields to enhance detection of small targets. For its detection head section, YOLOv7 employs three IDtect detection heads for different target sizes and utilizes Reparameterized Convolution (RepConv) to introduce learnable parameters into the convolution kernel. This adaptability allows for better capture of features in data and thus enhances model performance and generalization ability.

III. STRay Model

STRay is an improvement upon the YOLOv7 model. The enhanced model first utilizes the Swin Transformer attention mechanism at the end of the backbone network. It applies a sliding window multi-head self-attention mechanism to address the problem of low recognition rates caused by background interference and sample imbalance in complex backgrounds. Secondly, in the E-ELAN module, ordinary convolutions are replaced with deformable dilated convolutions. By learning sampling offsets to adapt to object deformations, the convolutional kernels better match the contours of prohibited items, effectively addressing the technical challenge of missed detection caused by mutual occlusion of prohibited items. Finally, the detection head of YOLOv7 is replaced with a more efficient decoupled head, decoupling the localization and classification tasks into separate feature channels, enhancing the classification and localization capabilities for small-sized prohibited items.

A. Swin Transformer Attention Mechanism

Due to overlapping and stacking of items in X-ray security inspection images, objects can occlude each other, and different materials result in a complex background with overlapping colors after passing through X-rays. This complexity makes it difficult for detection models to accurately identify and locate prohibited items. Therefore, this paper embeds the Swin Transformer attention mechanism at the end of the backbone network of the YOLOv7 model. Swin Transformer is a self-attention mechanism that replaces long sequences with a hierarchical sliding window approach, which can improve detection performance while minimizing the impact on runtime speed[12]. Traditional attention mechanisms dynamically emphasize regions of interest and suppress irrelevant background areas by learning weighted coefficients within the network. In contrast, self-attention calculates the relevance weights between features through matrix operations, enabling the model to capture relationships between features, which is suitable for detecting and identifying unclear features. The structure of the Swin Transformer attention mechanism is shown in (a) of Figure 2.

Firstly, the input is a three-channel feature image, which is segmented using the patch partition module. The segmented images are then fed into 4 stages for hierarchical attention computation. Except for stage 1, which uses a Linear Embedding layer, the remaining three stages downsample through Patch Merging. Within each stage, the Swin Transformer block serves as the core module for attention computation. It is constructed by repetitively stacking Window-based Multi-Head Self-Attention (W-MSA)[13] and Sliding Window-based Multi-Head Self-Attention (SW-MSA)[14]. In the Swin Transformer Block, as shown in the diagram on the right in (a), MLP represents Multi Layer Perceptron, LN represents Layer Normalization, and the output after each

module is represented by the following formula:

$$\hat{X}^l = \mathrm{W-MSA}\left(\mathrm{LN}\left(X^{l-1}\right)\right) + X^{-1} \tag{1}$$

$$X^l = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{X}^l\right)\right) + \hat{X}^l \tag{2}$$

$$\hat{X}^{l+1} = \mathrm{SW-MSA}\left(\mathrm{LN}\left(X^l\right)\right) + X^l \tag{3}$$

$$X^{l+1} = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{X}^{l+1}\right)\right) + \hat{X}^{l+1} \tag{4}$$

The internal mechanism of the Swin Transformer Block is illustrated in (b) of Figure 2. Based on the W-MSA, the feature image is divided into four windows labeled as A, B, C, and D. Since W-MSA only computes self-attention within each window, there is no information exchange between windows. To establish internal connections within the four windows, the feature image is further partitioned into nine windows labeled 1 to 9 using SW-MSA. This results in 2.25 times more computations for the 9 windows compared to the 4 windows. To ensure consistent window numbers for parallel MSA computation, a strategy involving upward and leftward shifts is employed to reorganize the 9 windows into 4 windows of equal size as A, B, C, and D. Finally, MSA computation is performed to output the feature map. Through subsequent ablation experiments, it was demonstrated that the use of the Swin Transformer enables the network to focus more on the recognition and localization of small prohibited items and effectively suppress background interference in complex backgrounds.

B. Deformable Dilated Convolution Module

In real-world X-ray security inspection images, contraband items are often randomly distributed at various locations in the image with diverse scales and different poses, leading to the problem of contraband items occluding each other in security inspection images. Regular CNN models with ordinary convolutions have fixed geometric structures, and the geometric structure of convolutional networks formed by their stacking is also fixed. To enhance the network's capability in recognizing objects with complex geometric deformations, this paper incorporates Deformable Convolution v2 within the E-ELAN module, replacing standard convolutions [15]. This approach introduces directional parameters to each element of the convolutional kernel, enabling it to dynamically adjust its shape over a wide range during training. This adaptability allows the kernel to better conform to the distinctive features of contraband items. The comparison between regular convolution and deformable convolution sampling points is shown in Figure 3. The deformable convolution depicted in the diagram not only shifts the input but also enables the adjustment of weights for each position input. The formula for adjustable deformable convolution can be expressed as follows:

$$y(p) = \sum_{k=1}^{\kappa} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \tag{5}$$

In the equation, Δp_k and Δm_k represent the learnable offset and adjustment parameters at the k-th position, where the adjustment parameter Δm_k is within the range [0,1], and

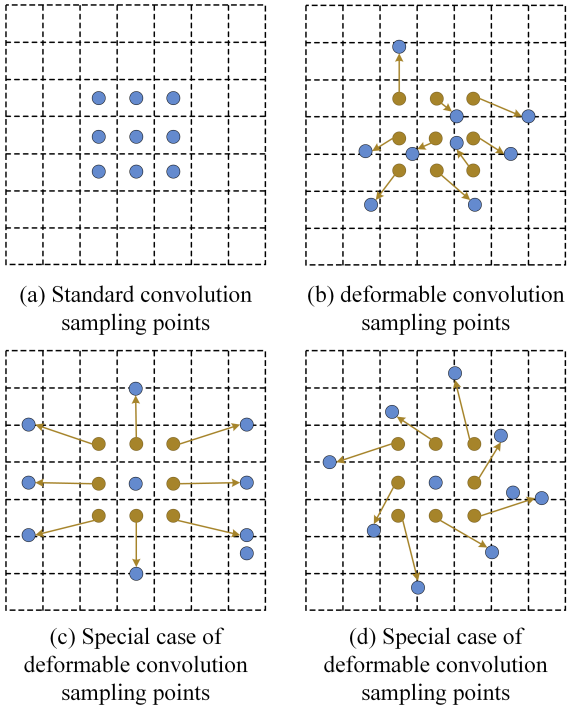

Fig. 3. Comparison between standard convolution and deformable convolution sampling.

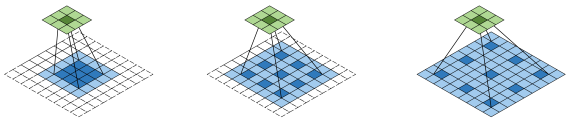

Fig. 4. Diagram of cavity convolution expansion rate.

Δp_k can take any value. To enhance the deformable convolution's ability to learn geometric transformations, dilated convolution is applied before using deformable convolution. Dilated convolution introduces a dilation rate parameter to define the spacing between convolution kernels, allowing for different receptive fields by setting different dilation rates[16]. Dilated convolution effectively expands the receptive field of output units at minimal computational cost, achieving this without enlarging the size of the convolution kernel. Sequentially stacking multiple dilated convolutions enhances its effectiveness in various applications. In Figure 4, (a) shows a 3×3 dilated convolution with a dilation rate of 1, resulting in a receptive field of 9; (b) demonstrates a 3×3 dilated convolution with a dilation rate of 2, leading to a receptive field of 25; (c) illustrates a 3×3 dilated convolution with a dilation rate of 4, resulting in a receptive field of 81. Thus, dilated convolution can enlarge the receptive field without affecting the image resolution.

Experimental verification has shown that applying deformable convolution on top of dilated convolution achieves two main objectives: on one hand, it makes the convolution kernel shape closer to the characteristics of prohibited items; on the other hand, it enlarges the receptive field, providing richer semantic information. By integrating these techniques, the model not only learns the overall contour but also gains more detailed information, effectively addressing the technical challenge of missed detection caused by prohibited

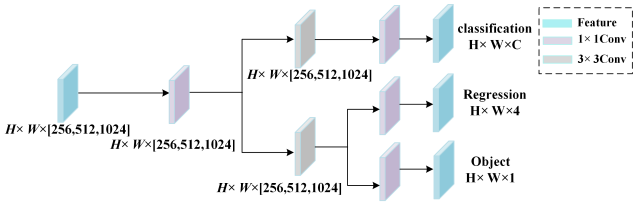

Fig. 5. Efficient Decoupled Detection Head Structure.

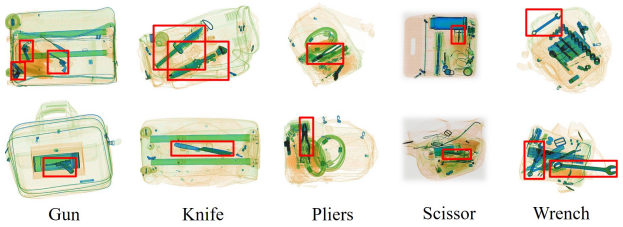

Fig. 6. Examples of prohibited items in the SIXray dataset.

items occluding each other.

C. Efficient Decoupled Detection Head

Because prohibited item detection involves multiple scales and objects, differences in scale can cause the detection model to overlook small targets, leading to missed detections. Therefore, accurate localization information and comprehensive classification information are essential for this task. To resolve these challenges, the paper introduces the Efficient Decoupled Detection Head (EDDH) to replace the original detection head for target prediction [17]. Initially, YOLOv7 employed a coupled detection head where parameters for classification and localization tasks were shared. However, this joint processing caused interference between classification and regression tasks, impacting detection accuracy. The EDDH separates these tasks, enabling the network to concentrate independently on each task. This separation enhances the model's accuracy, particularly in detecting small-sized targets. Figure 6 illustrates the structure of the Efficient Decoupled Head.

In Figure 5, the input feature data is subjected to channel adjustment using a 1×1 convolution. Subsequently, the feature map is fed into two parallel channels. Each channel consists of a 3×3 convolutional layer for extracting features. After feature extraction, the upper channel adjusts the number of feature channels to perform classification tasks. Meanwhile, the lower channel further splits into two sub-paths after feature extraction. One sub-path is responsible for determining bounding box parameters (height, width, and center coordinates), while the other sub-path focuses on obtaining confidence parameters. This approach enhances detection accuracy and improves network efficiency compared to traditional decoupled heads.

IV. DATASETS AND EVALUATION METRICS

A. Experimental Dataset

SIXray[18]: Developed by the Pattern Recognition and Intelligent Systems Development Laboratory at the University of Chinese Academy of Sciences, the dataset consists of 1,059,231 X-ray luggage images. Within this dataset, 8,929 images are annotated for object detection, focusing on 5 categories of prohibited items: guns, knives, wrenches, pliers, and scissors. For experimental purposes, these annotated images were randomly partitioned into training, testing, and validation sets in an 8:1:1 ratio.

OPIXray[19]: The dataset, constructed by Beihang University, comprises 8,885 X-ray security inspection images. It includes five categories of knives: Folding Knife(FK), Straight Knife(SK), Scissors(SC), Utility Knife(UK) and Multi-tool Knife(MK). In the experiments, the dataset was randomly divided into training, validation, and testing sets in a ratio of 8:1:1.

PIDray[20]: A large-scale security inspection image dataset constructed by the Chinese Academy of Sciences, comprising 47,677 images containing 12 categories of prohibited items. These categories include guns, knives, wrenches, pliers, scissors, hammers, handcuffs, batons, sprays, power banks, lighters, and bullets. The dataset consists of 29,457 training images and 18,220 test images (categorized by detection difficulty into 9,482 easy, 3,733 hard, and 5,005 heavily occluded). Due to the diverse range of prohibited items in this dataset, for this experiment, images of three detection difficulty levels—easy, hard, and hidden—are selected separately.

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:

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$AP = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$mAP = \frac{\sum_{n=1}^{Num(class)} AP(n)}{TP + TN + FP + FN} \tag{9}$$

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.

TABLE I
RESULTS OF COMPARATIVE EXPERIMENTS ON DIFFERENTAL GORITHMS BASED ON SIXRAY DATASET.

Algorithms	AP (%)					mAP (%)	FPS
	gun	knife	pliers	scissors	wrench		
Faster R-CNN	97.3	82.4	89.6	85.4	81.2	87.2	17.1
SSD	95.0	81.5	86.2	77.5	76.8	83.4	47.3
DenseNet	87.4	87.2	64.1	87.6	60.6	77.4	40.5
YOLOv3	95.3	79.9	78.1	73.5	74.1	80.2	54.0
YOLOv4	97.3	82.4	89.6	85.4	82.2	87.4	65.6
YOLOv5	98.4	85.2	95.2	85.6	87.5	90.4	78.2
YOLOv7	98.4	89.1	89.5	88.8	87.5	90.7	90.0
MTRay	**99.3**	**91.8**	**95.7**	**94.4**	**95.3**	**95.3**	**87.4**

TABLE II
RESULTS OF COMPARATIVE EXPERIMENTS ON DIFFERENTAL GORITHMS BASED ON OPIXRAY AND PIDRAY DATASET.

Algorithms	OPIXray						PIDray			
	FK	SK	SC	UK	MK	Average	easy	hard	hidden	Average
Fcos	86.4	68.5	90.2	78.4	86.6	82.0	61.8	51.7	37.5	50.3
SSD	76.9	35.0	93.4	65.9	83.4	70.9	68.1	58.9	45.7	57.6
YOLOv3	92.5	36.1	97.3	70.8	94.4	78.2	72.2	65.1	54.1	63.9
YOLOv5	92.0	65.2	97.9	74.1	93.2	84.5	78.9	73.4	68.1	73.5
YOLOv7	92.3	72.2	98.3	85.0	93.7	88.3	87.5	80.1	71.2	79.6
MTray	**92.7**	**75.7**	**98.4**	**82.5**	**94.9**	**88.8**	**91.0**	**85.4**	**72.9**	**83.1**

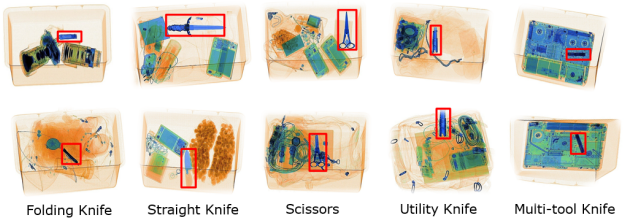

Fig. 7. Examples of prohibited items in the OPIXray dataset.

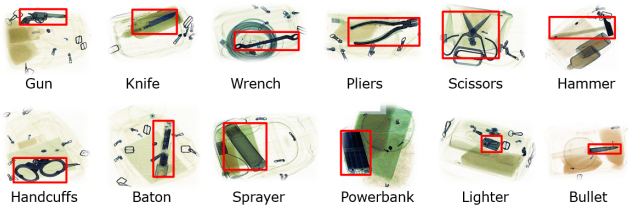

Fig. 8. Examples of prohibited items in the PIDray dataset.

V. Experiment and Result Analysis

This paper designs two types of experiments. The first type is comparative experiments, which verify the overall effectiveness of the proposed model by conducting experiments on three mainstream object detection models and the improved model proposed in this paper using the SIXray, OPIXray and PIDray datasets. The second type is ablation experiments, using experiments on the SIXray dataset as an example, where the improvement points are gradually added to the baseline model to validate the effectiveness of each module in the proposed model.

A. Experimental Configuration

The study was carried out using the Windows 10 operating system, utilizing PyTorch 1.12 and GPU NVIDIA RTX3080 for network framework development. The training utilized a batch size of 8 over 300 epochs. Stochastic Gradient Descent (SGD) was chosen for optimizing network parameters, starting with a learning rate of 0.01 and a weight decay coefficient set to 0.005. Additionally, the learning rate adjustments were made using the cosine annealing algorithm. For pre-training in our experiments, we utilized the original YOLOv7.pt weight file as it shares most of its structure with the improved model during training phase.

B. Contrast experiments

To verify the advancement of the improved model STRay proposed in this paper and its transferability to contraband X-ray images, experiments were conducted on three datasets: SIXray, OPIXray, and PIDray. Additionally, a comparative analysis was performed with the current mainstream object detection models: Faster R-CNN, SSD, Fcos[21], YOLOv3, YOLOv4, YOLOv5 and YOLOv7. The experimental results are presented in Tables 1 and 2.

To demonstrate the general applicability of the STRay model, this paper also conducted comparative experiments on the OPIXray and PIDray datasets. Due to the large number of categories of prohibited items in the PIDray dataset and its vast size, samples were divided into three categories: easy, hard, and hidden, for testing. As shown in Table 2, the detection accuracy of STRay on the OPIXray and PIDray datasets reached 88.8% and 83.1%, respectively, representing improvements of 0.5% and 3.5% compared to YOLOv7. Since all five categories of prohibited items in the OPIXray dataset are tools with similar shapes, and when the blade is placed vertically in the security screening machine, only a blue line is displayed on the X-ray security image, the improvement on the OPIXray dataset is not significant. STRay achieved good detection results on the large-scale PIDray dataset, with significant improvements in detecting prohibited items in simple, complex, and heavily obscured images. The PR curve plots of the YOLOv7 model and the improved STRay model on the SIXray, OPIXray, and PIDray datasets are shown in Figure 9. The top image

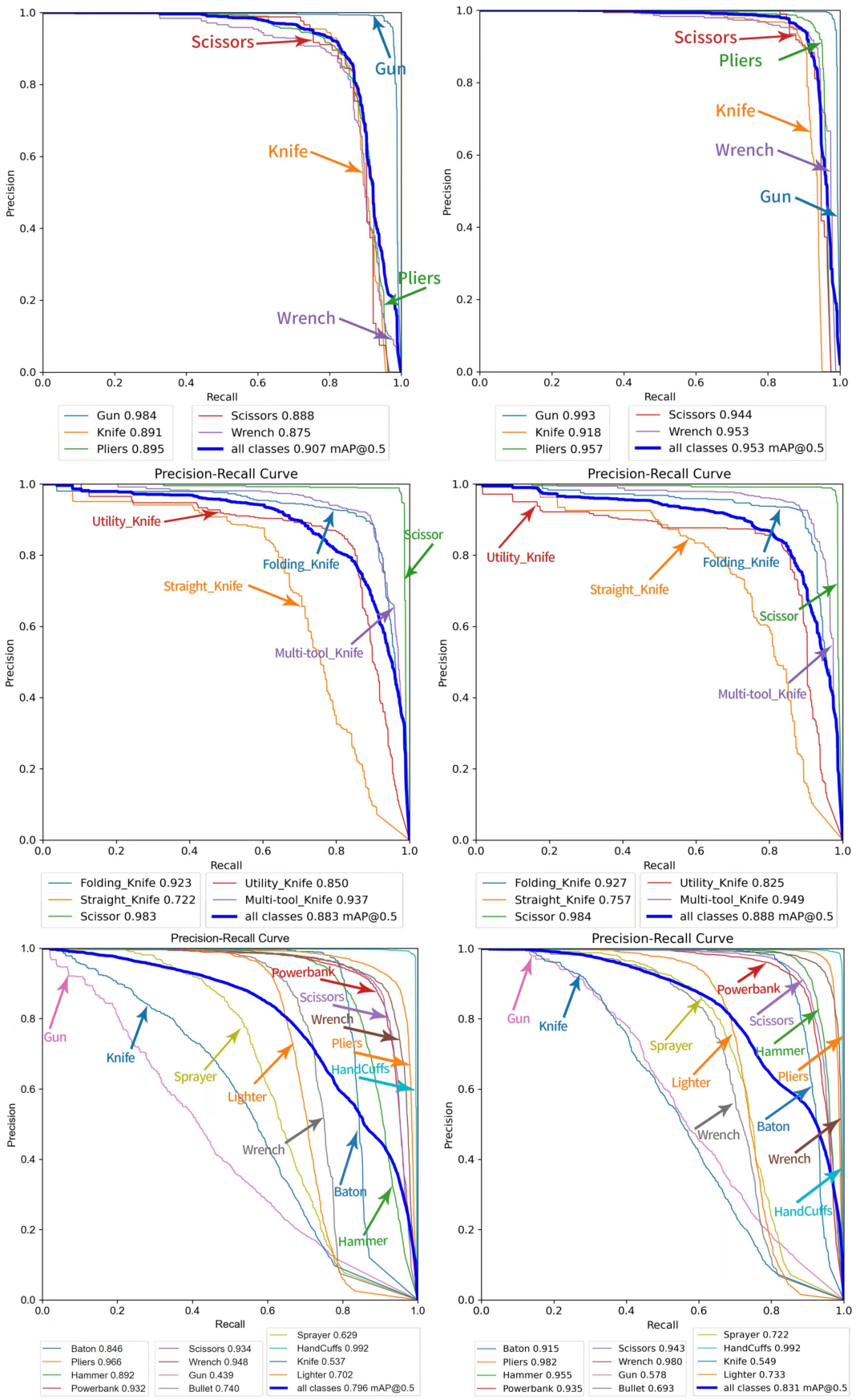

Fig. 9. Comparison of PR curves for YOLOv7 and MTRay on three datasets.

displays the PR curve of YOLOv7, while the bottom image shows the PR curve of the improved STRay proposed in this paper. It can be observed from the PR curve that the proposed STRay model provides more accurate identification

TABLE III
RESULTS OF ABLATION EXPERIMENTS ON THE SIXRAY DATASET.

Model	AP (%)					mAP (%)
	gun	knife	pliors	scissors	wrench	
YOLOv7	98.4	89.1	89.5	88.8	87.5	90.7
YOLOv7+A	98.9	90.7	92.5	91.9	90.4	92.9
YOLOv7+A+B	99.3	91.5	95.0	93.8	94.5	94.8
YOLOv7+A+B+C	99.3	91.8	95.7	94.4	95.3	95.3

of prohibited items in X-ray security images and exhibits good generalization and robustness.

C. Ablation Experiments

To validate the effectiveness of each improvement component on the STRay model, YOLOv7 was used as the baseline model, and ablation experiments were conducted by gradually incorporating each improvement module into YOLOv7. The experimental results are shown in Table 3. In the table, method A represents strengthening the feature extraction of contraband items by using Swin Transformer attention mechanism at the end of the backbone network, method B represents replacing ordinary convolutions with deformable dilated convolutions in E-ELAN, and method C represents replacing the detection head with a more efficient decoupled head. From the analysis of the table, it can be observed that the mAP improved by 2.2% after using Swin Transformer. Replacing ordinary convolutions with deformable dilated convolutions in E-ELAN increased mAP by 1.9%. Further, the mAP increased by an additional 0.5% after using the Efficient Decoupled detection head in the detection head section. Therefore, it can be concluded that the methods proposed in STRay can effectively improve the detection accuracy of contraband items.

VI. Conclusion

To address the issues of mutual occlusion of items and the presence of numerous small-sized contraband items in X-ray security inspection images, this paper proposes the STRay contraband detection model for X-ray security inspection images based on YOLOv7. STRay integrates Swin Transformer self-attention mechanism, deformable dilated convolution module, and Efficient decoupled detection head into a unified object detection model. Comparative experiments with other models demonstrate the advancement and robustness of the proposed model after improvements. In real-life scenarios, luggage carried by passengers is often cluttered, and there is a wide variety of contraband items with diverse shapes. Future work will focus on enhancing the model's ability to detect multi-scale contraband items and further improving the accuracy of contraband detection.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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, pp. 345–354, 2024.

[7] 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, Oct. 2022.

[8] F. Shao, J. Liu, P. Wu, Z. Yang, and Z. Wu, "Exploiting foreground and background separation for prohibited item detection in overlapping x-ray images," *Pattern Recognition*, vol. 122, p. 108261, Feb. 2022.

[9] B. Song, R. Li, X. Pan, X. Liu, and Y. Xu, "Improved yolov5 detection algorithm of contraband in x-ray security inspection image," in *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*. IEEE, Aug. 2022.

[10] J. Fu and Y. Tian, "Underwater target detection based on improved yolov7," *IAENG International Journal of Computer Science*, vol. 51, no. 4, pp. 422–429, 2024.

[11] B.-r. Li, J.-k. Zhang, and Y. Liang, "Pafpn-solo: A solo-based image instance segmentation algorithm," in *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*. IEEE, Mar. 2022.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021.

[13] C. Bai, F. Sun, J. Zhang, Y. Song, and S. Chen, "Rainformer: Features extraction balanced network for radar-based precipitation nowcasting," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, p. 1–5, 2022.

[14] Z. Zhou, S. Qiu, Y. Wang, M. Zhou, X. Chen, M. Hu, Q. Li, and Y. Lu, "Swin-spectral transformer for cholangiocarcinoma hyperspectral image segmentation," in *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, Oct. 2021.

[15] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017.

[16] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2018.

[17] K. Liu, Z. Lv, K. Xia, C. Zhou, Z. Lu, H. Zuo, Z. Li, and X. Chen, "Improved yolov5 based on deformable convolution and efficient decoupled head for pill surface defect detection," in *Third International Computing Imaging Conference (CITA 2023)*, X. Shao, Ed. SPIE, Nov. 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[19] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. ACM, Oct. 2020.

[20] B. Wang, L. Zhang, L. Wen, X. Liu, and Y. Wu, "Towards real-world prohibited item detection: A large-scale x-ray benchmark," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021.

[21] N. Wang, Y. Gao, H. Chen, P. Wang, Z. Tian, C. Shen, and Y. Zhang, "Nas-fcos: Fast neural architecture search for object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2020.