

SCD-YOLO: A security detection model for X-ray images based on the improved YOLOv5s

Xiaotong Kong, Aimin Li^(✉), Wenqiang Li, Zhiyao Li and Yuechen Zhang

¹ Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

² Shandong Engineering Research Center of Big Data Applied Technology, Faculty of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
lam@qlu.edu.cn

Abstract. X-ray security inspection is widely used in the subway, high-speed rail, airports, key locations, logistics, and other scenarios. However, because of the complexity and diversity of objects in the X-ray images in real-world scenarios, it is easy for security personnel to make mistakes or miss inspections when they are fatigued or not fully focused. In this paper, we proposed an improved model based on YOLOv5 to help security inspectors improve the efficiency of security inspection procedures. First, we replaced the SPP (spatial pyramid pooling) feature fusion module with SPPFCSPC to further enhance the feature extraction capability. Then, we added CoordConv before each feature map input to the detection head. This enables the model to perceive positional information and enhances its feature extraction capability, effectively addressing the detection of small prohibited items in complex backgrounds. Finally, we used decoupled detector head instead of the traditional coupled detector head to separate the classification and localization tasks further improves the detection speed. The experimental results show that our method achieves 77% accuracy. Compared with state-of-the-art methods, our model also achieves significant improvements in detection accuracy and recall.

Keywords: Security Object Detection, X-ray, YOLOv5s, Neural Network.

1 Introduction

1.1 A Subsection Sample

X-ray security inspection, as the primary means of security inspection, is widely used in the subway, high-speed rail, airports, key locations, logistics, and other scenarios. Due to the large inspection volume and the object's complexity, it is easy for security personnel to make mistakes or miss inspections when they are fatigued or not fully focused. On the field of X-ray image security detection, object detection algorithms have great potential and application prospects. By combining object detection algo-

rithms with X-ray images, we can achieve fast and accurate identification and positioning of potential threats and abnormal situations. This technology has a wide range of applications that cover aviation.

Object detection task is an important part of the field of computer vision. Before the popularity of deep learning, limited by computing resources and other constraints, traditional target detection usually focused on reducing the dependence on computing resources. Like Viola Jones Detectors (original slide windows algorithm) [1], HOG Detector (Histogram of directional gradients is used to describe features) [12], and Deformable Part-based Model (DPM) [3]. After deep neural networks gained attention, object detection can be divided into two categories: 'two-stage detection' and 'one-stage detection'. The usual method for two-stage models is first to generate a region proposal using feature extraction, and then to locate and classify objects based on the region proposal. Representative models include RCNN (Region-based Convolutional Neural Networks) [4], Fast RCNN [5], and Faster RCNN [6]. In 2015, Joseph proposed the groundbreaking detection model YOLO (You Only Look Once) [7,8,9]. The core idea of YOLO is to treat target detection as a regression problem and make predictions based on the whole image rather than region suggestions or sliding windows, which is fast and generalizes well. Since then, the YOLO series has been a focal research point for many scientists. After YOLOv5 was proposed, a large number of industrial applications emerged. Subsequent iterations like YOLOX [10], YOLOv7 [11], and other models with higher accuracy, such as the CornerNet [12] and ExtremeNet [13], also emerged. These detection models based on anchor-free [14] and DETR [15] series that appeared after the transformer [16] was introduced into the visual field; however, YOLOv5 is still the preferred choice in practical applications in terms of speed and accuracy balance.

Applying object detection algorithms to X-ray image security detection can not only improve detection efficiency and accuracy but also reduce the burden of manual operations, lower error rates, and demonstrate better application potential in some complex scenarios.

Due to the complexity and diversity of objects in X-ray security inspection images in real-world scenarios, as well as varying imaging angles, different levels of occlusion, and overlapping of multiple objects, issues such as missed detections, false detections, and multiple detections can easily occur. In order to make the model more likely to detect dangerous items and improve the accuracy of the model, we made the following specific improvements:

- Pyramid Pooling Module: The SPP (Spatial Pyramid Pooling) module [17] is replaced with the CSPP (Cross-Stage Partial Connection) structure [18] to enhance feature fusion and improve feature extraction capabilities. It effectively prevents the reduction of recall caused by overlapping objects that are difficult to detect.
- Coordinate Convolution: After each feature map is fed into the detection head, the CoordConv convolution is added to perceive spatial information better and improve spatial awareness.
- Decoupled head: The traditional coupled detection head is replaced with the decoupled head from YOLOX [10], which divides the classification task and

the positioning task into two independent tasks, further improving the detection speed and accuracy.

2 Related Work

In recent years, X-ray image security detection technology has developed rapidly, aiming to improve the ability to identify hidden or blocked contraband in complex security inspection scenarios. A key research focus is on the development of effective methods for removing occlusion effects in images and improving the accuracy and reliability of detection systems. In 2019, a pioneering work by Miao et al. laid the foundation for this field [19]. The de-occlusion attention module (Depth Attention Module), designed by the researchers, employed deep learning technology to enhance attention to the characteristics of items obscured by occlusion, resulting in a notable enhancement in detection performance. In conjunction with this innovation, the researchers also constructed and publicly released the Occlusion Prohibited Items X-ray (OPIXray) data set, which represents the inaugural high-quality benchmark data set in the domain of security inspections. This data set has significantly contributed to the advancement and assessment of related algorithms. In 2021, Tao et al. advanced the state of the art by creating the HiXray security inspection image dataset [20], which markedly enhanced both quality and diversity. Building upon this foundation, they introduced the lateral inhibition module (Longitudinal Inhibition Module, LIM). This design is inspired by the way the human visual system processes overlapping object information. It suppresses irrelevant information and focuses on analyzing key identifiable features, thus maintaining efficient recognition in complex situations where objects cover each other. Tao et al. (2022) [21] focused on endogenous shift, where the differences between domains are mainly caused by intrinsic factors (e.g., imaging mechanisms, hardware components, etc.) and are usually inconspicuous. Then, they contribute the first Endogenous Domain Shift (EDS) benchmark, X-ray security inspection. Liu et al. (2023) [22] have proposed adversarial attacks that are valuable for evaluating the robustness of deep learning models. They develop a differentiable converter that facilitates the generation of 3D-printable objects with adversarial shapes, using the gradients of a surrogate model rather than directly generating adversarial textures. Furthermore, they present the physical-world X-ray adversarial attack dataset XAD, providing a valuable resource for evaluating and enhancing the attack resistance of existing detection models.

3 Method

3.1 SCD-YOLO

YOLOv5 achieves competitive accuracy in object detection tasks, especially in detecting small objects and crowded scenes. However, due to the inherent characteristics of the single-stage target detection model, the detection speed is faster, but the accuracy is slightly lower than the two-stage object detection model. In particular, the accuracy

of small object and crowded scene detection needs to be improved. In response to the above problems, we made three improvements to the original model.

Firstly, to improve the feature extraction capability of the model, we replaced the original feature pyramidal grouping method of the model with SPPFCSPC [23]. Enhance the model's ability to capture global and contextual information while expanding the receptive field. Secondly, we introduce coordinate convolution(CoordConv) [24] to enable our model to better perceive the position information of objects in the image. Finally, replace the detection head with decoupled head. Each task head is responsible for handling a specific task. This improves resource utilization efficiency while reducing the number of model parameters and inference speed and accuracy. The improved model structure is illustrated in Fig. 1.

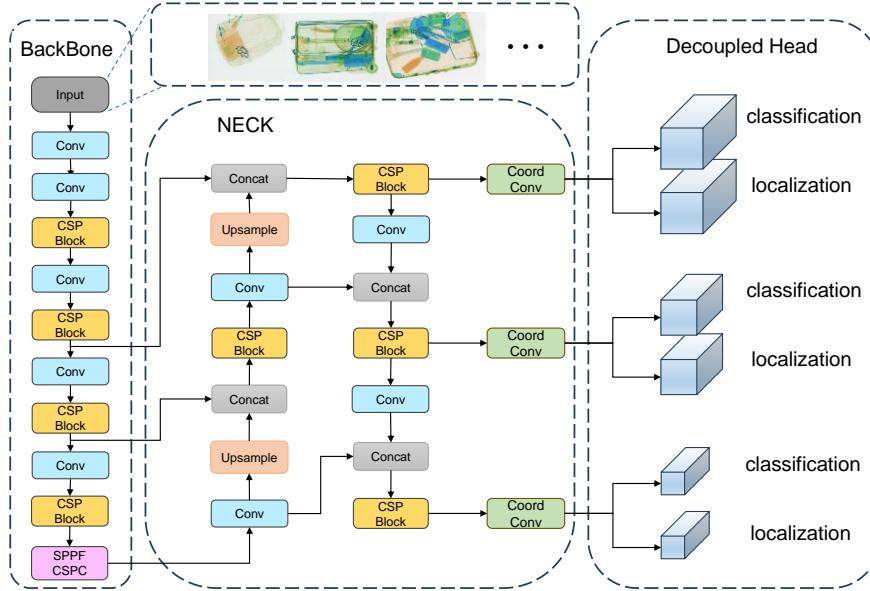


Fig. 1. The architecture of our model.

3.2 SPPFCSPC

Based on the innovative SPPCSPC module in YOLOv7, it has been noted that it enhances the model's ability to adapt to scale variations by employing different sizes of max-pooling layers to capture different sizes of receptive fields. The experimental results demonstrate that this approach leads to significant performance improvements. To enhance the effectiveness of the YOLOv5 model in multi-scale feature fusion, we plan to incorporate the CSPC (Cross Stage Partial Connections) structure into the existing SPPF (Spatial Pyramid Pooling Fusion) module of YOLOv5.

A new module called SPPFCSPC was designed, which contains two parallel branches. One of the branches will be directly involved in the final feature stitching

process, while maintaining the original simplicity and efficiency. The other branch undergoes two 1x1 convolutional layers and one 3x3 convolutional layer before entering the pooling phase. This aims to achieve deep fusion and dimensionality reduction of the input features, in order to extract more representative high-level features. The main innovation lies in the adoption of the idea of pooling branches with different convolutional kernel sizes (5x5, 9x9, 13x13) from the SPPCSPC module. This idea has been converted into a single branch, but with the implementation of three consecutive max-pooling operations of 5x5 in this branch. This is done to simulate the different scales of receptive fields covered by the original three branches. The feature maps generated by each max-pooling covered by the original three branches. The feature maps generated by each max-pooling covered operation will be used in the preliminary feature linking session to improve the speed of the model operation without sacrificing the original receptive field. The structure is depicted in Fig. 2.

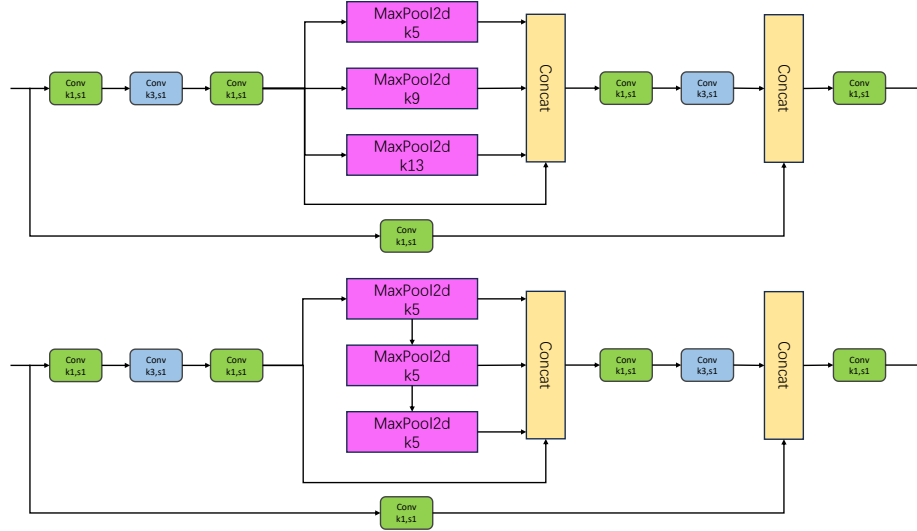


Fig. 2. Structural comparison of SPPCSPC(top) and SPPFCSPC(bottom).

3.3 CoordConv

X-ray security inspection images often have overlapping objects. Traditional convolutions have translation invariance, which allows them to learn essential features for tasks such as classification. However, when positional information needs to be perceived, the limitations of traditional convolutions become apparent. In order to allow convolutions to perceive spatial information, we introduced CoordConv to replace some of the traditional convolutions, and a CoordConv is added after each output feature map to further improve the performance. The main principle is to add two coordinate channels (representing the x and y coordinates of the original input) behind the input feature map and then perform traditional convolutions. This enables the convolution process to perceive the spatial information of the feature map, and this method is called CoordConv.

Using CoordConv, the network can learn translation invariance or a certain degree of translation dependency based on different task requirements.

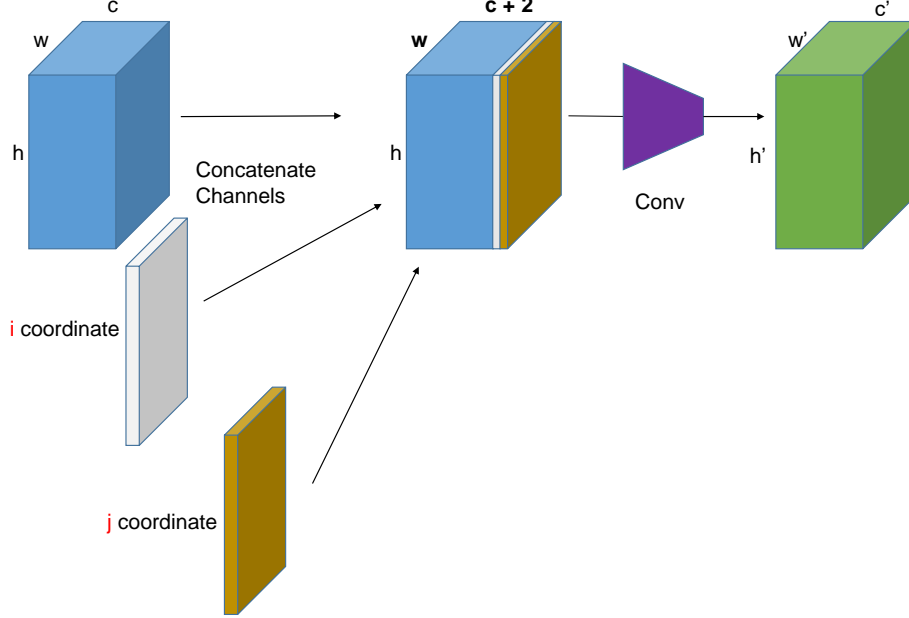


Fig. 3. The structure of CoordConv.

Fig. 3 shows the operation where two coordinates, i and j , are added. Specifically, the i coordinate channel is an $h \times w$ rank-1 matrix with its first row filled with 0's, its second row filled with 1's, its third row filled with 2's, and so on. The j coordinate channel is similar, but the columns are filled with constant values instead of rows. And use a final linear scaling to both the i and j coordinate values so that they fall in the range $[-1,1]$ For convolution over two dimensions, two (i, j) coordinates are sufficient to fully specify an input pixel, but, if desired, additional channels can be added to bias the models towards learning particular solutions. It can also use a third channel for an r coordinate, where

$$r = \sqrt{(i - h/2)^2 + (j - w/2)^2} \quad (1)$$

3.4 Decoupled Head

Typically, conventional detector head structures process the extracted high-level feature maps directly. They use convolutional or fully-connected layers to output the object's positional coordinates (bounding box) and category information. In contrast, the decoupled head adopts a more refined and targeted structural design.

The decoupled head utilizes a 1×1 convolutional layer to decrease the number of channels in the feature map to 256. This aims to refine and condense the key features

and reduce the complexity of subsequent computation. Following this, the architecture establishes two parallel branch networks, each containing two 3x3 convolutional layers. These two branches are divided into two tasks. The first is dedicated to the classification task, which involves extracting rich category information from the features and precisely determining which category the object belongs to. The second branch focuses on the regression task, which involves pinpointing the exact position of the object in the image, i.e. the bounding box coordinates. The reason for adopting a decoupling strategy is that object classification and location positioning require different feature understanding and parsing, and focus on different information dimensions. Separating them into independent branches helps the model to focus on mining the key features required by each, which is expected to improve overall detection performance.

In addition, the design of decoupled heads helps to reduce the number of parameters and computational complexity of the model. This, in turn, reduces the risk of model overfitting and enhances the model's generalization ability and robustness in different scenarios. The modular structure design not only improves algorithm execution efficiency but also shows higher accuracy and adaptability in practical applications. The structure as shown in Fig. 4.

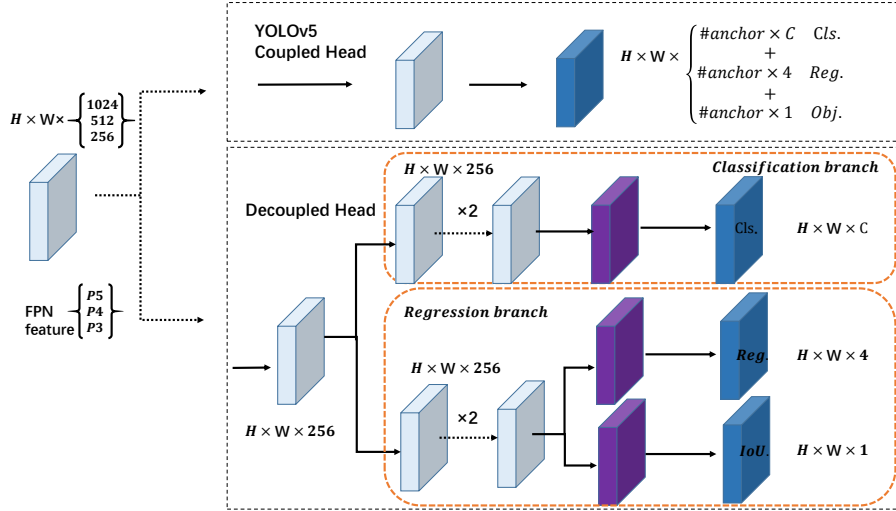


Fig. 4. The structure of decoupled head.

4 Experimental Results and Discussion

4.1 Dataset Source and Composition

All the images in this experiment are from the EDS [21] dataset. The EDS data set consists of domain1, domain2 and domain3, each containing approximately 5000 images. The images were captured by different detection machines, resulting in a total of 14,219 images. The dataset is divided into 10 categories: knife, glass bottle, pressure,

scissor, umbrella, power bank, lighter, laptop, device, and plastic bottle. To increase the amount of data, we merged the three parts of the dataset and divided them into training, validation, and testing sets in a ratio of 8:1:1.

4.2 Experimental Configuration

The training for this experiment was conducted on an NVIDIA GeForce RTX 2080ti GPU with 11GB memory. The system used was Centos7 and the experimental framework was Python-3.7.16 torch-1.12.1, with CUDA version 10.2. The batch size was set to 16, and the number of epochs was set to 200. For inference, a Tesla T4 GPU with 16GB memory was used and the batch size was set to 16. All other configurations remained consistent with the training environment.

4.3 Comparison Experiments

To validate the advantages of our proposed SCD-YOLO model over existing mainstream object detection models on the X-ray contraband detection task, we conducted an exhaustive comparative review on the EDS dataset. We selected several benchmark models, including RT-DETR [25], YOLOX [10], YOLOv8n, and YOLOv8s, and compared their accuracy and average precision mAP50 metrics for each contraband category. The results are presented in Table 1.

Table 1. Comparison of classification average accuracy AP (%), mean average accuracy mAP50 (%) of proposed SCD-YOLO, YOLOv5s, YOLOX, YOLOv8n, YOLOv8s and RT-DETR.

	knife	glassbottle	scissor	umbrella	pressure	laptop
YOLOv5s	59.3	68.3	53.7	94.2	86.4	84.8
RT-DETR	60.6	67.8	57.6	93.6	84.9	83.1
YOLOX	58.7	69.4	49.9	94.4	86.3	83.4
YOLOv8n	57.6	70.0	54.0	95.3	85.7	82.6
YOLOv8s	64.4	72.7	58.2	96.3	89.5	86.0
ours	64.9	73.1	58.3	96.4	89.9	87.3

Table 1. Comparison of classification average accuracy AP (%), mean average accuracy mAP50 (%) of proposed SCD-YOLO, YOLOv5s, YOLOX, YOLOv8n, YOLOv8s and RT-DETR (continued).

	powerbank	device	lighter	plasticbottle	mAP50 (%)
YOLOv5s	70.9	79.0	71.2	73.1	74.1
RT-DETR	71.5	75.1	69.6	73.5	73.7
YOLOX	64.4	75.3	66.5	69.9	71.8
YOLOv8n	62.5	75.0	65.9	67.6	71.6
YOLOv8s	69.2	80.1	71.2	73.4	76.1
ours	72.8	81.4	72.7	73.6	77.0

The table data clearly shows that our SCD-YOLO model outperforms the original YOLOv5 and other comparative models in terms of overall performance and in each specific category. Notably, the detection accuracy for three categories of contraband, namely knives, glass bottles, and pressure vessels, has significantly improved compared to YOLOv5s, with an increase of 5.6%, 4.8%, and 4.6% in accuracy, respectively.

On the mAP50 metric, which is a comprehensive measure of detection performance, the SCD-YOLO model improved by 3.3% compared to RT-DETR, 5.2% compared to YOLOX, 5.4% compared to the YOLOv8n version, and 0.9% compared to the YOLOv8s, albeit with a smaller overall improvement. Finally, an increase of 2.9% compared to the original YOLOv5 model. These results demonstrate the superiority of our model in the field of X-ray contraband detection and the effectiveness of the improvements.

4.4 Ablation Experiments

In order to further verify the validity of the three modules added to the original YOLOv5 model and the feasibility of the SCD-YOLO model, we conducted ablation experiments on the test dataset for the three innovations improved in this paper. It can be seen from Table 2. Ablation Experiments. that each of our improved modules shows varying degrees of improvement in the mAP50, mAP50:95 and recall metrics compared to the original YOLOv5 model. This suggests that our modifications are effective in improving the detection accuracy of the model. Firstly, we evaluate the mAP50, mAP50:95 and Recall metrics of the original YOLOv5 model, which yield results of 74.1%, 50.4% and 66.8% respectively. Then we introduced the SPPFCSPC module, and this improvement resulted in the model’s mAP50, mAP50:95 and Recall improving by 0.6%, 0.9% and 0.8% relative to the original model. After introducing the Coord-Conv, mAP50, mAP50:95 and Recall achieved 66.9%, 74.3%, and 51% respectively. In addition, by adding the Decoupled head, mAP50, mAP50:95 and Recall improved 1%, 1.2% and 1.5% respectively. Finally, our model SCD-YOLO improved the mAP50 metrics by 2.9%, mAP50:95 by 1.6%, and the Recall metrics by 3% over the original YOLOv5 model.

Table 2. Ablation Experiments.

SPPFCSPC	Coord-Conv	Decoupled head	mAP50 (%)	mAP50:5:95 (%)	Recall (%)
			74.1	50.4	66.8
✓			74.7	51.3	67.6
	✓		74.3	51.0	66.9
		✓	75.1	51.6	68.3
✓	✓		75.0	50.5	69.0
✓		✓	75.3	52.2	69.6
	✓	✓	75.4	52.1	68.7
✓	✓	✓	77.0(+2.9)	52.0(+1.6)	69.8(+3.0)

We also compared the P-R curve between the two models. As shown in Fig. 5, it is evident that our improved model outperforms the original model. This shows that the improved model achieved more ideal results in balancing precision and recall, significantly improving the model's superior ability to identify target categories.

Fig. 5. P-R training curves of the YOLOv5s(left) model and ours(right).

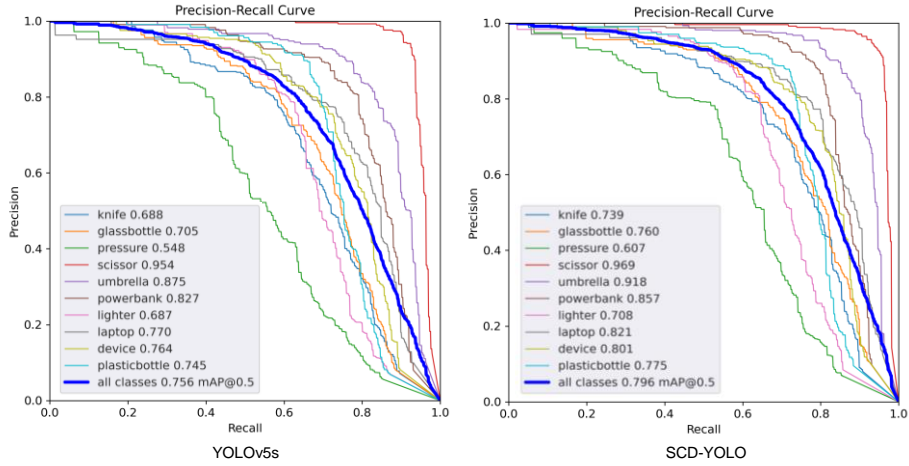


Fig. 5. P-R training curves of the YOLOv5s(left) model and ours(right).

The experimental results comprehensively demonstrate the effectiveness of our innovation in the object detection task, affirming the contribution of these methods to performance enhancement. Our approach performs admirably across various evaluation metrics, thereby further validating its potential for practical applications.

Finally, we present several sets of comparison images to briefly examine the actual detection results in Fig. 6.

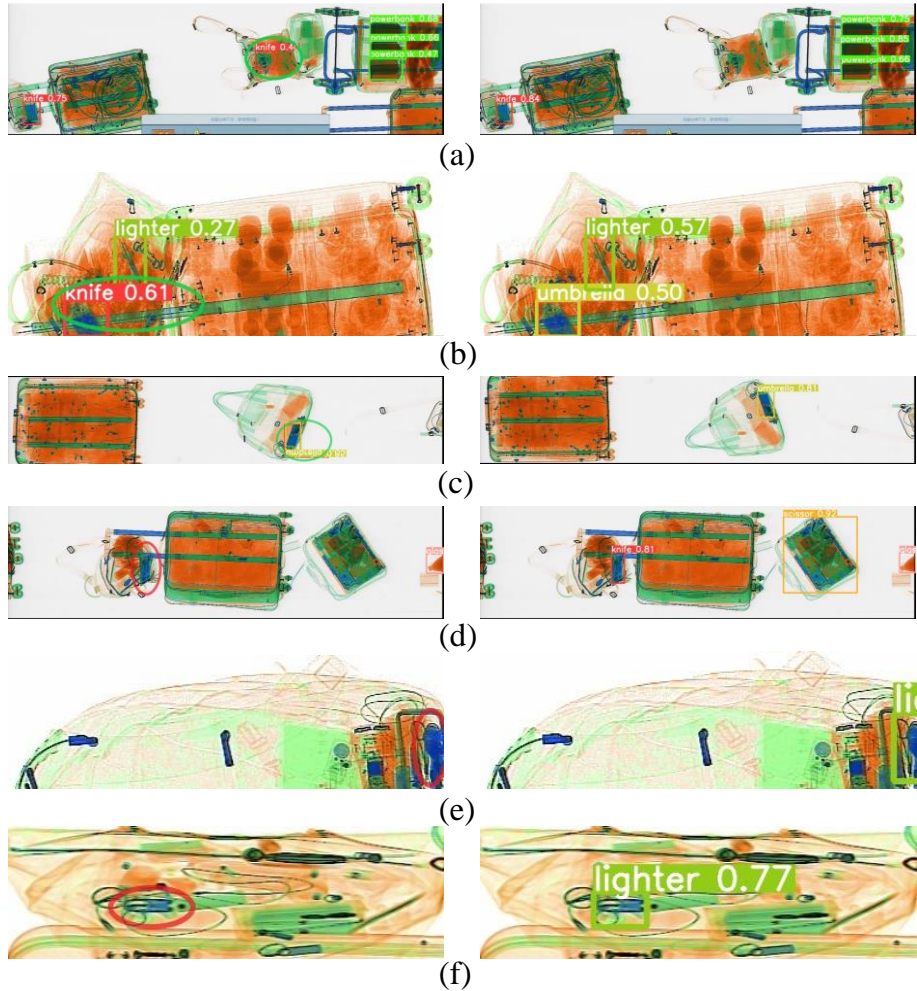


Fig. 6. Actual detection results comparison: on the left side is yolov5s, and on the right side is the improved model.

In the first three sets of images, (a)(b)(c) represents that the original model often has false detections for dangerous items like knives. In the fourth set of images (d), the original model missed the detection of a knife, and in the fifth and sixth sets (e)(f), it missed the detection of a lighter. The improved model shows significant improvements in these cases. This indicates that the improved model performs better in detecting dangerous items, with a reduction in the number of false alarms and missed detections. Our method enhances target perception and improves positional accuracy, thereby increasing overall robustness.

5 Conclusion

In this paper, we proposed a security detection model for X-ray security inspection images to improve the accuracy of contraband detection. The experimental data showed that our improvement model SCD-YOLO have significantly improved accuracy and recall compared to YOLOv5 and other mainstream detection models, which means that our model can detect more contraband with higher accuracy under the same circumstances. While this model may not completely replace human work and there are areas that necessitate further improvement, it has shown significant promise and practical significance.

Acknowledgments. This work was supported by the Key R&D Plan of Shandong Province, China (No.2021CXGC010102).

References

1. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, pp. I-I. Ieee, (2001)
2. Navneet, D.: Histograms of oriented gradients for human detection. In: International Conference on Computer Vision & Pattern Recognition, 2005, pp. 886-893. (2005)
3. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: 2008 IEEE conference on computer vision and pattern recognition, pp. 1-8. Ieee, (2008)
4. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587. (2013)
5. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. (2015)
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, pp. 91-99. MIT Press, Montreal, Canada (2015)
7. Redmon, J., Farhadi, A.J.a.p.a.: Yolov3: An incremental improvement. (2018)
8. Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger. In: IEEE Conference on Computer Vision & Pattern Recognition, pp. 6517-6525. (2016)
9. Redmon, J., Farhadi, A.J.a.e.-p.: YOLOv3: An Incremental Improvement. (2018)
10. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding YOLO Series in 2021. (2021)
11. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.J.a.e.-p.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. (2022)
12. Law, H., Deng, J.: CornerNet: Detecting Objects as Paired Keypoints. Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XIV, pp. 765–781. Springer-Verlag, Munich, Germany (2018)
13. Zhou, X., Zhuo, J., Krhenbühl, P.: Bottom-up Object Detection by Grouping Extreme and Center Points. (2019)
14. Huang, L., Yang, Y., Deng, Y., Yu, Y.J.C.S.: DenseBox: Unifying Landmark Localization with End to End Object Detection. (2015)

15. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, pp. 213–229. Springer-Verlag, Glasgow, United Kingdom (2020)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.J.a.: Attention Is All You Need. (2017)
17. Kaiming, Zhang, Xiangyu, Shaoqing, Jian: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. (2015)
18. Wang, C.Y., Liao, H.Y.M., Yeh, I.H., Wu, Y.H., Chen, P.Y., Hsieh, J.W.: CSPNet: A New Backbone that can Enhance Learning Capability of CNN. (2019)
19. Miao, C., Su, C., Wan, F., Liu, H., Jiao, J., Xie, L., Ye, Q.: SIXray : A Large-scale Security Inspection X-ray Benchmark for Prohibited Item Discovery in Overlapping Images. (2019)
20. Tao, R., Wei, Y., Jiang, X., Li, H., Qin, H., Wang, J., Ma, Y., Zhang, L., Liu, X.: Towards Real-world X-ray Security Inspection: A High-Quality Benchmark and Lateral Inhibition Module for Prohibited Items Detection. (2021)
21. Tao, R., Li, H., Wang, T., Wei, Y., Ding, Y., Jin, B., Zhi, H., Liu, X., Liu, A.: Exploring Endogenous Shift for Cross-domain Detection: A Large-scale Benchmark and Perturbation Suppression Network. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21157-21167. IEEE Computer Society (2022)
22. Liu, A., Guo, J., Wang, J., Liang, S., Tao, R., Zhou, W., Liu, C., Liu, X., Tao, D.: X-Adv: physical adversarial object attacks against X-ray prohibited item detection. *Proceedings of the 32nd USENIX Conference on Security Symposium*, pp. Article 212. USENIX Association, Anaheim, CA, USA (2023)
23. Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., Li, Y., Zhang, B., Liang, Y., Zhou, L., Xu, X., Chu, X., Wei, X., Wei, X.J.A.: YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. [abs/2209.02976](https://arxiv.org/abs/2209.02976), (2022)
24. Liu, R., Lehman, J., Molino, P., Such, F.P., Frank, E., Sergeev, A., Yosinski, J.: An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution. (2018)
25. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.J.a.p.a.: Detsr beat yolos on real-time object detection. (2023)