Real-time detection network of contraband based on YOLOX

Shifeng Jiang, Chunrong Pan, Yufeng Gan, Junjie Chen, Xiangdong Gao*
School of Mechanical and Electrical Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, Jiangxi, China

ABSTRACT

Today, the increasingly complex external environment and the goal of building a harmonious society have put forward higher requirements for security check work. Traditional security inspection methods are facing enormous challenges. With the maturity of intelligent security technology, Intelligent security inspection has become a trend. Security check that requires both reasoning speed and accuracy, this paper proposes a lightweight target detection network, which takes YOLOX as the overall framework and Mobilenet-V3 as the backbone network. The effectiveness of the network is proved by the simulation experiment on the public data set OPIXray and can achieve a real-time computing speed of 87.7FPS.

Keywords: Contraband detection, deep learning, YOLOX, real-time detection

1. INTRODUCTION

Nowadays, with the rapid development of the economy and the continuous improvement of transportation infrastructure, high-speed rail, airplanes, and other travel methods are becoming ever more popular, so the challenges to ensure the safety of public travel have become more severe. As a widely used security screening technology, the X-ray security inspection of luggage that based on the different X-ray absorption rates of the items, and then the security personnel checks the images to determine whether there are prohibited items. However, heavy and tedious detection work makes the accuracy of manual detection easy to be disturbed by many factors. In order to improve detection efficiency, automatic detection of prohibited items can be used as an auxiliary means.

In the past, the automatic detection of prohibited items mainly used hand-designed features[1, 2], such as scale-invariant feature transform (SIFT) and Gabor texture features[3-5]. However, the complex environment inside the luggage and the perspective effect of X-ray imaging make the stacking of objects in the image serious, so traditional machine learning methods are difficult to identify accurately. In recent years, with the great success of deep learning in computer vision, the use of deep convolutional neural networks for the automatic detection of contraband targets has become a mainstream approach. To address the occlusion problem of security screening images, Wei et al.[6] apply the attention mechanism to the input image and propose a de-occlusion attention module that uses the edge and material information of the contraband to generate an attention map, effectively improving the detection accuracy. Mu et al.[7] based on the YOLOv4 feature fusion network, added an atrous dense convolutional layer to expand the receptive field and used an attention module to filter essential features before the prediction network.

It is undeniable that the existing CNN-based models have achieved good performance in contraband detection, but there is a common problem that the model is large in scale. Now, the actual Baggage security devices do not have enough computing power resources to run large models. Want to deploy deep learning models on devices with limited computing power and resources, building lightweight networks is an efficient approach. For the target detection network, lightweight and high precision are mutually opposed. Pursuing the light weight of the network will inevitably result in loss of accuracy, and high precision often requires a large model as support. In order to achieve a balance between lightness and accuracy, this paper selects YOLOX as the network framework to study the performance of the lightweight backbone network on X-ray images.

2. METHOD

2.1 YOLOX

In the target detection task, the YOLO series is a classic and effective one-stage target detection algorithm that can achieve

* 1427422015@qq.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250619
 · 0277-786X · doi: 10.1117/12.2662518

real-time and end-to-end detection. The YOLO family of algorithms has been updated with multiple versions, such as YOLOv5, and YOLOX[8] is an anchor-free target detector based on YOLOv5. Like other networks in the YOLO series, YOLOX continues the basic idea of dividing the image into NXN grids, and the grid is responsible for predicting objects. However, unlike YOLOv5, which uses the anchor template to predict the width and height of the detection box. YOLOX directly predicts the width and height information, which alleviates the problems of sample imbalance and numerous hyperparameters (anchor frame size, quantity, and aspect ratio) caused by the dense setting of the anchor template. Without the participation of the anchor template, the sample matching method of calculating the IOU between the ground-truth box and the anchor template is no longer applicable. YOLOX defines the cost function to measure the matching degree of the network prediction result and the ground-truth box, and models the sample matching as an optimal transmission problem. In the design of the detection head, YOLOX uses decoupled head instead of a YOLO head to alleviate the spatial misalignment in the two subtasks of localization and classification, and speed up the convergence of the network.

2.2 MobileNetV3

MobileNetV3[9] is a lightweight convolutional neural network that inherits and improves the inverted residual structure in MobileNetV2. The original inverted residual structure uses depthwise separable convolution to extract features after projecting feature maps to high dimensions, then maps back to low dimensions and performs linear activation. The inverted bottleneck structure effectively improves memory efficiency, and the depthwise separable convolution can significantly reduce the number of network parameters and computational costs with a slight decrease in accuracy. In order to improve performance, MobileNetV3 adds a SE attention module to the inverted residual structure to model the relative importance of channels, dynamically adjusts the response of each feature map to suppress invalid feature maps and strengthen key feature maps. In addition to this, a new nonlinear activation function h-swish is introduced in the deep layer of the network. Compared with the swish activation function, h-swish has a faster calculation speed and has no difference in accuracy with swish. After the above-mentioned overall structure is determined, the model hyperparameters are optimized by the neural architecture search technology, and the network performance is further enhanced.

2.3 YOLOX_Mobilenetv3

This paper builds a one-stage target detector based on YOLOX and MobileNetV3 to realize the automatic detection of contraband targets. The overall architecture of the network is shown in Figure 1. In YOLOX, although CSP-Darknet53 can reduce the computational load and memory usage, it is more complex than the lightweight network and needs a lot of computing resources. To reduce the complexity of the network structure, MobileNetV3 is used as the new backbone network. The original MobileNetV3 is designed for image classification tasks, so the feature map output has a low resolution. To obtain a feature map suitable for contraband detection, the last global average pooling layer and the fully connected layer of MobileNetV3 are discarded, select the feature maps with down sampling 8, 16, and 32 times for prediction. The feature fusion module and decoupling detection head of the network refer to the structural design in YOLOX_s, which further reduces the number of parameters and computational costs. The parameters of each backbone network and model are shown in Table 1. It can be seen that compared with YOLOX_s, the target detector using MobileNetV3 as the backbone network has fewer parameters and less computation. Finally, in the calculation of regression loss, CIOUloss is used instead of IOUloss to speed up network convergence and achieve higher localization accuracy.

3. EXPERIMENT AND RESULT

3.1 Data set

Compared with natural images, X-ray images are more difficult and expensive to obtain, so the number of security inspection X-ray image datasets is small. The public datasets include GDXray, SIXray, and OPIXray. In this paper, the OPIXray dataset is selected for the experiment, which is designed for object detection tasks in security screening scenarios. OPIxray dataset has a total of 8885 images of prohibited knives, including straight knives, scissors, folding knives, utility knives, and multi-purpose knives, among which 7,109 images constitute the training set, and the remaining 1,776 images are used for testing. The 1776 test images are divided into three test sets according to the degree of occlusion. The data set is shown in Table 2.

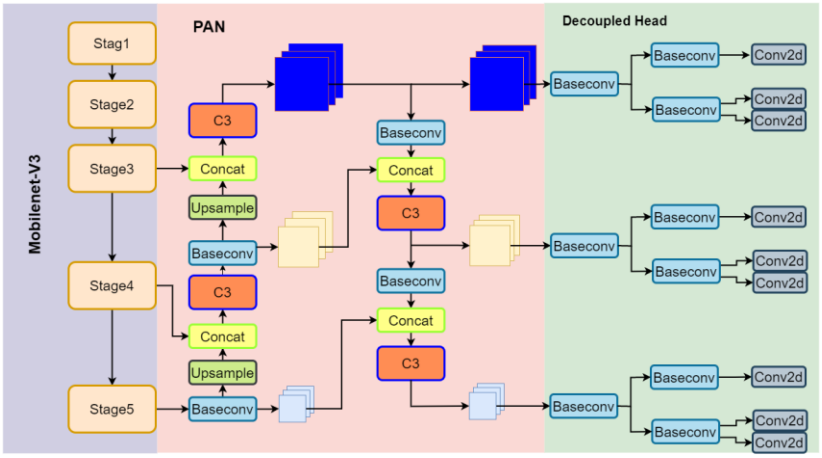

Figure 1. Diagram of the YOLOX_Mobilenetv3.

Table 1. Parameters of backbone network and object detection model.

Network	Params/M	FLOPS/G
CSP-Darknet53	27.1	76.1
MobileNet-v3	3.0	3.8
YOLOX	54.2	155.7
YOLOX_s	8.9	26.8
Ous	8.2	20.1

Table 2. OPIXray data set.

OPIXray	Categories					Total
	Folding	Straight	Scissor	Utility	Multi-tool	
Train	1589	809	1494	1635	1612	7109
Test	404	235	369	343	430	1776
Total	1993	1044	1863	1978	2042	8885

3.2 Result and analysis

In this paper, mAP and FPS are selected as evaluation indicators to evaluate the accuracy and inference speed of the network. During the experiment, choose YOLOX_s for comparative experiments. As can be seen from Table 3, in the complete test set, the mAP of YOLOX_s is 89.93%, while the mAP of the target detector using MobileNetV3 as the backbone network has reached 91.83%, and the AP of various targets has improved, straight knife increased the most, by 7.4%. More importantly, the FPS of the MobileNetV3-based object detector is 15% higher than that of YOLOX_s, reaching 87.7FPS, which strongly indicates that MobileNet-V3 can better balance the relationship between lightness and accuracy. To further verify the effectiveness of the model, we conduct comparative experiments on test datasets with different occlusion degrees. The experimental results are shown in Table 4. As the degree of occlusion increases, the accuracy of the network decreases, but the MobileNetV3-based object detector is still better than YOLOX_s. It is also worth noting that the straight knife has the lowest AP of all classes detected by both models. By visualizing the detection results (as shown in Figure 2), it can be observed that since some straight knife shapes are flat and have a similar color to the background when placed vertically, they become a line in the image and are difficult to identify. The same happens with other prohibited knives and the object detector is prone to false detection objects with similar shapes.

Table 3. Comparison of detection results of target detection models.

Method	mAP	FPS	Folding AP	Straight AP	Scissor AP	Utility AP	Multi-tool AP
YOLOX_s	89.93	76.2	92.99	75.96	97.93	88.89	93.87
Ous	91.83	87.7	94.12	83.37	98.21	89.09	94.34

Table 4. The mAP of model under different occlusion conditions test set.

Method	OL1	OL2	OL3
YOLOX_s	90.11	90.22	88.47
Ous	92.31	91.02	90.68

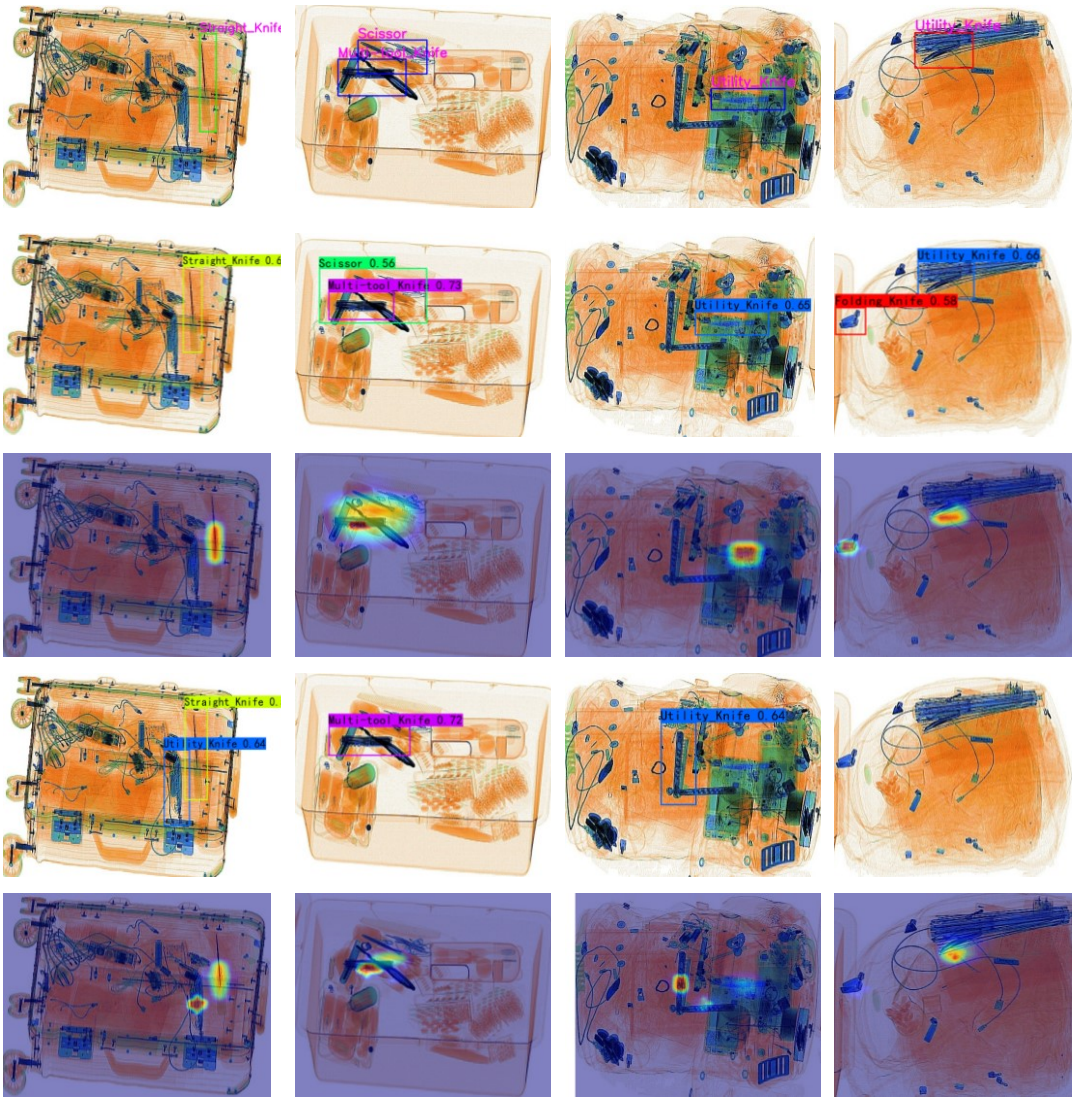

Figure 2. prediction results visualization and heatmap: (a) target ground truth; (b) YOLOX_Mobilenetv3 detection results and heatmap; (c) YOLOX_s detection results and heatmap.

4. CONCLUSION

In this paper, we use YOLOX as the main framework and Mobilenetv3 as the backbone network to build a lightweight one-stage target detector. Through comparative experiments on the OPIXray dataset, we analyze the influence of the lightweight backbone network on the model performance. The experimental results show that using Mobilenetv3 as the backbone network of the target detector can speed up the inference speed while ensuring accuracy. Similarly, the experiment also revealed some problems, randomly placed prohibited objects and objects of a similar shape easily being ignored and false detection. In addition, due to the overlapping and perspective phenomena in X-ray images, it is difficult to detect small objects. Next, we will introduce Transformer[10] to solve these problems in the future.

REFERENCES

[1] Chen, J. Q., Wang, T., Gao, X. D. and Li, W., "Real-time monitoring of high-power disk laser welding based on support vector machine," Computers in Industry, 94, 75-81(2018).

[2] Zhang, Y. X., Han, S. W., Cheon, J., Na, S. J. and Gao, X. D., "Effect of joint gap on bead formation in laser butt welding of stainless steel," Journal of Materials Processing Technology, 249, 274-284(2017).

[3] Wang, T., Gao, X. D., Seiji, K. Y. and Jin, X. L., "Study of dynamic features of surface plasma in high-power disk laser welding," Plasma Science and Technology, 14(3), 245(2012).

[4] Gao, X. D., Ding, D. K., Bai, T. X. and Katayama, S., "Weld-pool image centroid algorithm for seam-tracking vision model in arc-welding process," IET Image Processing, 5(5), 410-419(2011).

[5] Gao, X. D., and Zhang, Y., "Monitoring of welding status by molten pool morphology during high-power disk laser welding," Optik-International Journal for Light and Electron Optics, 126(19), 1797-1802(2015).

[6] Wei, Y., Tao, R., Wu, Z., Ma, Y., Zhang, L. and Liu, X., "Occluded prohibited items detection: An X-ray security inspection benchmark and de-occlusion attention module," Proceedings of the 28th ACM International Conference on Multimedia, 138-146(2020).

[7] Mu, S. Q., Lin, J. J., Wang, H. Q. and Wei, X. Z., "X-ray image contraband detection algorithm based on improved YOLOv4," Acta Armamentarii, 42(12), 2675(2022).

[8] Ge, Z., Liu, S., Wang, F., Li, Z. and Sun, J., "Yolox: Exceeding yolo series in 2021," arXiv preprint arXiv:2107.08430, (2021).

[9] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M. and Adam, H., "Searching for mobilenetv3," Proceedings of the IEEE/CVF International Conference on Computer Vision, 1314-1324(2019).

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. and Polosukhin, I., "Attention is all you need," Advances in Neural Information Processing Systems, 30, (2017).