PIDNet: Prohibited Items Detection Network and Fine-Coarse Encoder Module

Yu Yao[2], Boliang Zhang[2(✉)], H. K. Kan[1], and Chan Tong Lam[2]

[1] Macao Polytechnic University, Centre for Continuing Education, Macao, China
hykan@mpu.edu.mo
[2] Macao Polytechnic University, Faculty of Applied Sciences, Macao, China
P1807471@mpu.edu.mo

Abstract. Security inspection using X-rays are absolutely familiar in everyday life and have an essential function in protecting public safety. However, it is not straightforward to perceive the presence of prohibited items, and the key challenge is that any prohibited items in X-ray images may exhibit color-monotonous and luster-insufficient, mainly due to the characteristics of X-ray imaging mechanisms. In this paper, to address this problem, we constructed a fresh prohibited items detection dataset (PIDD) and proposed a prohibited items detection network (PIDNet), which searches enrichment fine-grained and coarse-grained features for powerful prohibited items detection with a novel Fine-Coarse Encoder (FCE) module. Extensive experiment demonstrates that our proposed method achieves significantly superior contraband detection results on the PIDD test set compared to progressive methods for prohibited items detection, effectively proving the practicability of the method proposed in this paper.

Keyword: Prohibited Items · Environmental Security · Security Inspection · X-ray Images

1 Introduction

Airport public environmental security is an eternal topic, which is most inseparable from the silent payment of security inspectors, as well as the detection of dangerous goods security machine. However, human security screeners are susceptible to exhaustion after long hours of highly concentrated work, which in turn leads to a reduced ability to recognise prohibited items, which may pose a serious risk to the public. Therefore, the development of a fast, accurate, and automated assisted detection algorithm is a necessity.

With the rapid development of deep learning techniques in the field of computer vision [4, 5], especially convolutional neural networks, this can be achieved by converting them to the task of object detection in computer vision. Unfortunately, the X-ray images generated by the security machine are colour-monotonous and luster-insufficient and existing detection algorithms cannot be effectively applied directly to this task [6–8].

Specifically, when we examine the existing object detection models, the traditional CNN-based method [9–12] model the target features by performing the extraction of the

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 574, pp. 279–290, 2024.
https://doi.org/10.1007/978-3-031-65123-6_20

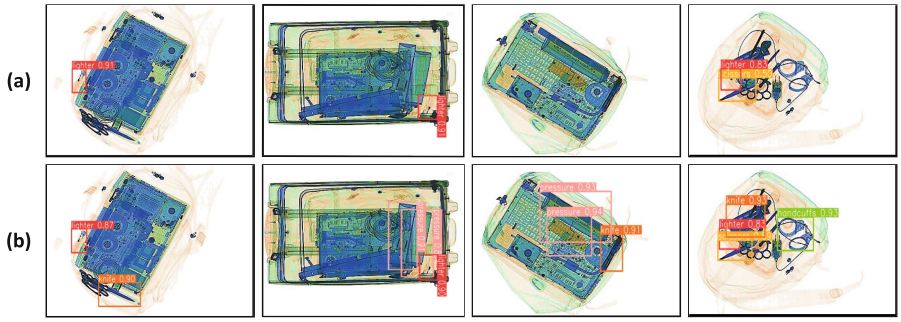

Fig. 1. Comparison of detection results before and after improvement. Existing object detection methods, including efficient You Only Look Once v8 (YOLOv8) [1], are trained on datasets that are biased towards rich texture information. While these methods work well in good conditions [2, 3], they fail in rare texture events (top). The scarcity of these features is a challenge for basic object detection methods. The proposed method (bottom) learns in colour-monotonous and luster-insufficient training data and learns targets that are otherwise invisible to basic detection algorithms.

target features and leads to missed inspections (as shown in (a) of Fig. 1) in this case because, unlike natural images with explicit texture information [13, 14], X-ray images [15–17] are precisely characterised by the lack of strong recognition properties and contain more noise. This urgently requires researchers to make targeted improvements to the model.

In order to solve the above problems, many researchers have proposed a range of algorithms for detecting X-ray prohibited items based on CNN models. However, a deep learning method with strong generalisation capabilities is extremely dependent on a large and diverse dataset. For a long time, only the public dataset GDXray [18] was available and only 1,552 X-ray baggage images were labelled. To overcome this dilemma, researchers have used techniques such as data augmentation and transfer learning to expand the dataset. Jain et al. [19] in order to expand the dataset used a generative model to generate new X-ray images. Cui and Oztan [20] used the Dangerous Goods Image Projection (TIP) generate algorithms to create a portion of the data and use it for training data. At the same time, a number of other studies have been devoted to augmenting the training data with generative data [21–23]. However, in further studies, the researchers found that the model trained with synthetic data has significant limitations in detecting prohibited items in X-ray security images and does not directly improve the training of the network [22, 24]. At this point, the researchers have once again begun to study the problem of prohibited item detection with the construction of a new dataset [7, 25].

In summary, we found that for a long time, researchers have been in the "data volume panic", constantly trying to tackle the difficult issue of prohibited items object detection in X-ray images from the data perspective. This phenomenon also leads to the fact that the essential problem of target detection in X-ray images has not been effectively solved. This intrinsic problem can be attributed to the high noise level of X-ray images, the lack of texture information, and the lack of clarity of structures due to the frequent overlapping of objects in the images.

In this work, we propose a Coarse-Fine Encoder module for prohibited items detection in Security inspection X-rays image. Specifically, for the problem of colour-monotonous and luster-insufficient X-ray images, this paper proposes the use of Coarse Encoder and Fine Encoder feature extractors with densely connected network structure to extract the overall textual information and to generate globally-associated coarse- and fine-grained features, respectively. The feature extractor models the global textual information using the contextual association structure and makes full use of the contextual association information to generate coarse-grained and fine-grained features, effectively overcoming the colour-monotonous and luster-insufficient problems of X-ray images (The improved detection effect is shown in Fig. 1(b)).

Under this method, the differences and correlations between coarse-grained and fine-grained features are fully considered, and the problem of target detection of multi-scale features is taken into account, which can effectively carry out the detection work applicable to the task. Our contributions are:

1) This paper proposes a plug-and-play Coarse-Fine Encoder module for the lack of texture information in X-ray images. The module can be directly applied to the existing target detection network, and effectively improve the detection accuracy of prohibited items in X-ray images.
2) We integrate and produce a new X-ray prohibited items detection dataset, PIDD, consisting of 10,605 images.
3) We conduct extensive experiments on the PID dataset integrated in this work and the results demonstrate the effectiveness of our method.

2 Related Work

X-ray imaging plays a crucial role in medical image analysis [26, 27] and security inspection [25, 28]. However, the fact that X-ray images are difficult to obtain has slowed down the progress of security detection research in computer vision due to the lack of dedicated high-quality datasets. Several recent works [7, 18] started to work on building such datasets. For example, the released benchmark dataset GDXray [18] contains 19,407 greyscale images, a portion of which contains three prohibited items including guns, short swords, and razor blades. SIXray [26] is a large X-ray dataset, 100 times larger than the GDXray dataset, but in order to simulate a similar test environment, it has less than 1% positive samples, and the classifications are done with classification labels annotated at the time of annotation. Recently presented the OPIXray dataset containing 8885 X-ray images from 5 classes of tools. Unfortunately, the images in the OPIXray dataset are synthetic, which directly affects model training. Other related works [29, 30] do not provide data downloads.

However, the above research only seeks to address the problem of prohibited items detection from the perspective of dataset richness. These studies have focused on the construction of datasets, but have neglected to explore the nature of the prohibited items detection task.

Therefore, this paper investigates the colour-monotonous and luster-insufficient problem of prohibited items in X-ray images, and proposes a new Fine-Coarse Encoder (FCE) module, whose multi-level feature encoding method effectively improves the detection accuracy of prohibited items.

3 Method

We discovery that the human visual system can recognise the presence of prohibited items well from the perspective of coarse-grained and fine-grained features (e.g., structural relationships between different objects) by considering texture-less but geometrically well-structured objects. This inspired us to utilise clear geometric structural relationships consisting of rich coarse-grained and fine-grained features to detect prohibited items.

For this purpose, we propose a novel Fine-Coarse Encoder (FCE) module to extract rich coarse-grained and fine-grained features from a large field for context inference and prohibited items localization. FCE module is designed to effectively integrate multi-scale large-field coarse-grained and fine-grained features for detecting prohibited items of various scale. Our proposed FCE module is applied in the Prohibited Items Detection Network (PIDNet), which aims to obtain different levels of coarse and fine grained features by multi-level multiplexing of the same multiple features for robust prohibited items detection in various scenarios.

3.1 Overview of the Network Structure

Figure 2 presents the proposed prohibited items detection network (PIDNet). The core idea of PIDNet is to perform multi-level feature extraction on images with sparse original features and to process the different levels of features in sub-channels. We propose a Fine-Coarse Encoder (FCE) module to separate and learn coarse and fine grained features at different levels.

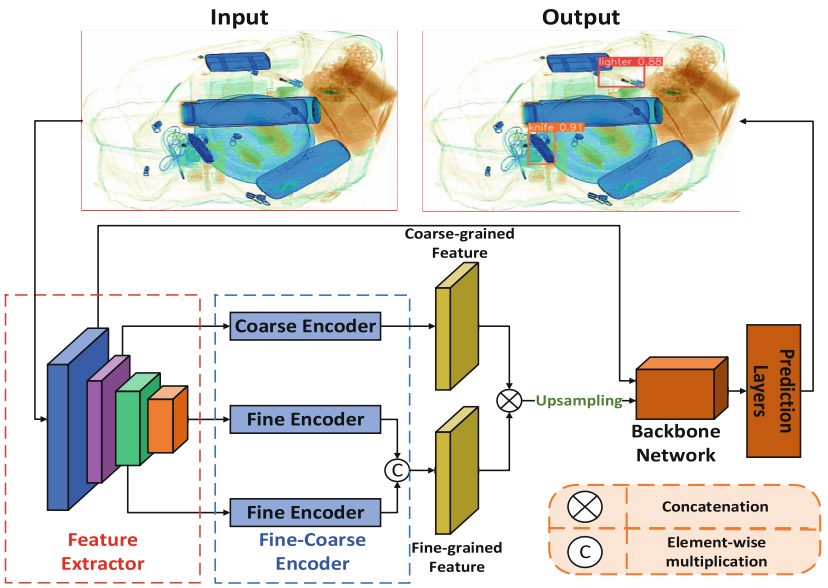

Fig. 2. The structure of the proposed PIDNet. We employ the pre-trained ResNeXt101 [31] as a Feature Extractor to acquire features at different levels. Fine-Coarse Encoder (FCE) is a plug-and-play module that we propose mainly for obtaining fine-grained and coarse-grained feature information at different hierarchical multi-levels.

The well-designed FCE module is capable of deep extraction of otherwise sparse and inconspicuous feature information, thus achieving the purpose of enhancing the prominence of features.

3.2 Fine-Coarse Encoder

Figure 2 illustrates the framework of our FCE module. Given the input features, the FCE module aims to efficiently and effectively extract and integrate multi-scale fine-grained and coarse-grained features, for the present of detecting prohibited items of various scale. FCE aims to productively extract rich coarse-grained and fine-grained features information from large fields for prohibited items detection.

FCE consists of two FE and one CE, Fine Encoders (FE) is used to extract high-level fine-grained features, Coarse Encoder (CE) is used to extract low-level coarse-grained features. With any FE or CE consisting of seven Residual-Dense Encoder blocks.

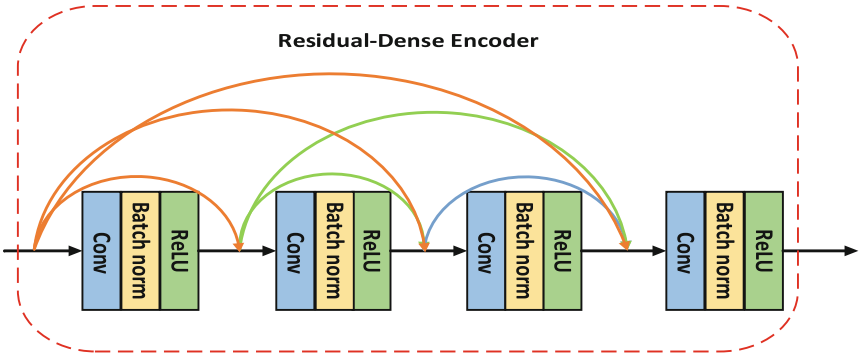

Fig. 3. Residual-Dense Encoder network architecture. The first three layers of the structure are in the form of a Residual-Dense structure, which aims at reusing features and avoiding feature extinction during the transfer process. The last layer is used to normalise and fine-tune the extracted features.

As shown in Fig. 3, each Residual-Dense Encoder blocks consists of four convolutional blocks. The first three of these layers will take all previous layers concate as input:

$$x_l = H_l([x_0, x_1, \cdots, x_{l-1}]) \tag{1}$$

where l denotes the number of layers in which the network is located. H represents the computational process carried out by the entire convolutional block.

The last layer is used to fine-tune the extracted features:

$$x_4 = H_4(H_3([x_1, x_2, x_3])). \tag{2}$$

3.3 Loss Function

We optimise the network parameters during training using three loss functions: binary cross entropy (BCE) loss L_{bce}, edge loss L_{edge} [32] and IoU loss L_{iou} [33].

Specifically, for high-level fine-grained feature parameter training, we combine the L_{bce} and the L_{iou} with the aim of forcing them to explore high-level feature information used complete prohibited items detection. For low-level coarse-grained features, we expect them to provide more geometrically clear low-level feature information. Therefore, we combine the L_{bce} with the L_{edge}, which will induce the network parameters to be more inclined to retain clear geometrically structured relationships belonging to the prohibited items.

For the final overall output, we optimise by combining the three losses, i.e., $L_f = L_{bce} + L_{iou} + L_{edge}$, with the aim of guiding the network to have a global knowledge of the prohibited items features.

Finally, the overall loss function is:

$$Loss = \lambda_{fine} L_{fine} + \lambda_{coarse} L_{coarse} + \lambda_f L_f \tag{3}$$

where λ_{fine}, λ_{coarse} and λ_f denote the balancing parameters for L_{fine}, L_{coarse} and L_f, respectively.

4 Experiment

4.1 Datasets Details

To better study the prohibited items detection problem, we collated and constructed a new large-scale contraband detection dataset (PIDD).We integrated public data on the internet to construct this dataset. It contains 10,605 pairs of prohibited items and prohibited items mask images. The dataset contains ten different types of prohibited items and the histogram of their class distribution is shown in Fig. 4, which clearly shows our dataset. For training and testing, we randomly divide 70% of the dataset into a training set and the other 30% into a testing set.

4.2 Experiments Settings

Implement Details. We implemented the model proposed in this paper using the PyTorch framework and used two GeForce RTX 3090Ti for training and testing the model. We used the SGD optimiser to optimise the model parameters. We trained the model for 1000 epochs and activated linear learning rate decay after 800 epochs.

Metrics. In traditional object detection tasks, performance metrics commonly used to measure system performance include mean Average Precision (mAP), Precision and Recall. This paper also use these performance metrics to judge the detection capability of the model, which are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

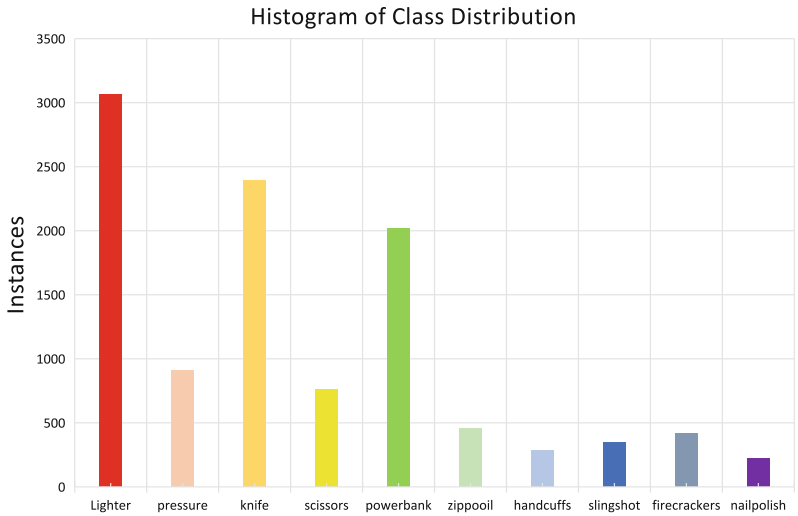

Fig. 4. Histogram of class distribution. From the above figure, it can be seen that the most numerous class in this dataset is Lighter, while the less numerous classes are zippooil, handcuffs, slingshot, firecrackers and nailpolish.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

where TP is a positive sample predicted as positive by the model, FN is a positive sample predicted as negative by the model, and TN is a negative sample that is predicted to be negative by the model.

The IOU is a measure of the degree of overlap between two targets (for target detection), mathematically described as:

$$IOU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{6}$$

where B_{gt} denote the Ground Truth (GT) of the target and B_p denote the predicted frame.

AP is to calculate the area under the $P - R$ curve of a certain type, and mAP is to calculate the average of the area under the $P - R$ curve of all types.

$F_1 score$ is defined as the harmonic average of precision and recall.

$$F_1 score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{7}$$

Baselines. We compare the prohibited items detection method proposed in this paper with five state-of-the-art target detection methods. Specifically, the competing methods are: YOLOv6 [27], YOLOv7 [7], YOLOv8 [36], SSD [52], FCOS [54]. We used the publicly available code for the above models and the parameter settings recommended in the code, and all models were retrained on the PIDD training set as a fair comparison test.

4.3 Experiments Result and Analysis

As described in Sect. 3, PIDNet performs well in the detection of prohibited items in X-ray images. We investigate the performance of five strategies on test set. Table 1 reports quantitative results of four metrics on test set. Obviously, it is shown that PIDNet achieves higher values.

Table 1. The results of the comparisons with other model. This table shows the performance metrics comparing the detection results of the proposed detection algorithm in this paper and other detection algorithms.

Method	Precision ↑	Recall ↑	mAP50 ↑	mAP50–95 ↑
FCOS	0.7248	0.3530	0.4012	0.2852
SDD	0.7610	0.3428	0.3927	0.2845
YOLOv6	0.8082	0.3266	0.3795	0.2772
YOLOv7	0.8392	0.3500	0.4062	0.3195
YOLOv8	0.8168	0.3443	0.4034	0.3079
PIDNet	0.8542	0.35352	0.4228	0.3311

Figure 5 further shows some prohibited items detection results on PID test set. We can find that PIDNet shows good detection results in these different scenarios, further proving the effectiveness of PIDNet.

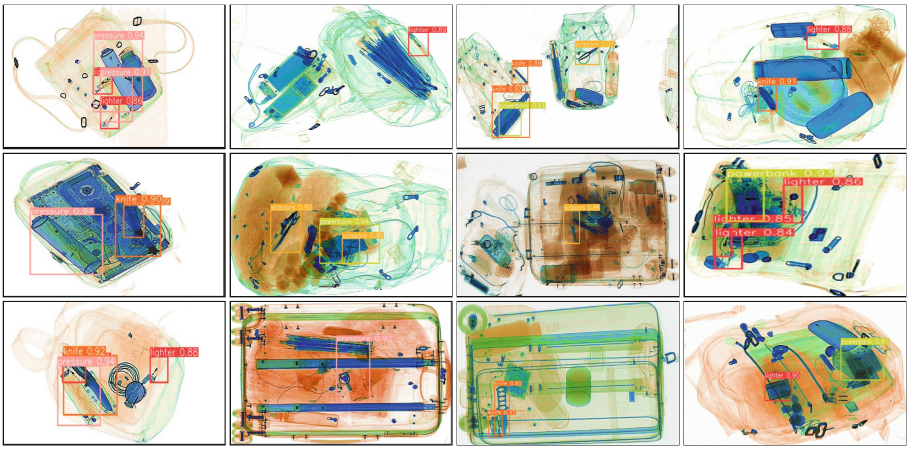

Fig. 5. Prohibited items detection results. The above figure shows the detection results of PIDNet proposed in this article. We can clearly see that PIDNet performs well in various scenarios and has good detection performance for detection objects of different scales.

Furthermore, we validated the effectiveness of this mechanism by applying FCE to five object detection methods, including traditional SSD [10], FCOS [34] YOLOv6 [35],

YOLOv7 [36], and YOLOv8 [1]. Table 2 shows the results of the ablation experiments performed on the PIDD dataset for the FCE module proposed in this paper. We can see a certain improvement in all four metrics when our method is added to different basic object detection methods.

Table 2. Component analysis. This table shows the results of ablation experiments on the proposed FCE module. "+FCE" means to add the FCE module proposed in this paper to the base model.

Method	Precision ↑	Recall ↑	mAP50 ↑	mAP50–95 ↑
FCOS	0.7248	0.3530	0.4012	0.2852
FCOS + FCE	0.7650	0.3396	0.3868	0.2851
SDD	0.7610	0.3428	0.3927	0.2845
SDD + FCE	0.7678	0.3108	0.3519	0.2442
YOLOv6	0.8082	0.3266	0.3795	0.2772
YOLOv6 + FCE	0.8272	0.3540	0.4220	0.3337
YOLOv7	0.8392	0.3500	0.4062	0.3195
YOLOv7 + FCE	0.8513	0.3497	0.4031	0.3098
YOLOv8	0.8168	0.3443	0.4034	0.3079
YOLOv8 + FCE	0.8542	0.3535	0.4228	0.3311

Based on the experimental data in Table 1 and Table 2 we can verify that our proposed model clearly has the best performance in the dimensions of the four indicators.Meanwhile, in the ablation experiments, it is again confirmed that our FCE module can effectively improve the effectiveness of the prohibited items detection model, which further proves its effectiveness.

5 Conclusion

In this paper, we provide a Prohibited Items Detection Dataset (PIDD) that is large-scale and contains multiple detection scenarios. We also propose a new network (PIDNet) to better solve the prohibited item detection task. PIDNet is able to detect prohibited items of different sizes in various scenarios using fine- and coarse-grained features extracted from different scales of feature information. We demonstrated the effectiveness of our network by performing an extensive evaluation validation on PIDD test set images. In addition, our method may fail in cases where the scene is very complex or where prohibited items and other objects are stacked on top of each other. We hope to better address this issue in the next phase of our research work.

Acknowledgments. This work is supported in part by the research grant (No.: RP/ESCA-07/2021) offered by Macao Polytechnic University.

References

1. Aboah, A., Wang, B., Bagci, U., et al.: Real-time multi-class helmet violation detection using few-shot data sampling technique and yolov8. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5349–5357 (2023)
2. Terven, J., Cordova-Esparza, D.: A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond. arXiv preprint arXiv:2304.00501 (2023)
3. Ju, R.Y., Cai, W.: Fracture Detection in Pediatric Wrist Trauma X-ray Images Using YOLOv8 Algorithm. arXiv preprint arXiv:2304.05071 (2023)
4. Qiu, J., Yan, X., Wang, W., et al.: Skeleton-based abnormal behavior detection using secure partitioned convolutional neural network model. IEEE J. Biomed. Health Inform. **26**(12), 5829–5840 (2021)
5. Wang, W., Yu, X., Fang, B., et al.: Cross-modality LGE-CMR segmentation using image-to-image translation based data augmentation. IEEE/ACM Trans. Comput. Biol. Bioinf. **20**, 2367–2375 (2022)
6. Tao, R., Wei, Y., Jiang, X., et al.: Towards real-world X-ray security inspection: a high-quality benchmark and lateral inhibition module for prohibited items detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10923–10932 (2021)
7. Wei, Y., et al.: 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, pp. 138–146 (2020)
8. Liu, A., Guo, J., Wang, J., et al.: X-adv: physical adversarial object attacks against x-ray prohibited item detection. arXiv preprint arXiv:2302.09491 (2023)
9. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
10. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision–ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
11. Wu, D., Lv, S., Jiang, M., et al.: Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. Comput. Electron. Agric. **178**, 105742 (2020)
12. Mei, H., Yang, X., Wang, Y., et al.: Don't hit me! glass detection in real-world scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3687–3696 (2020)
13. Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S.: Focusing attention: towards accurate text recognition in natural images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5076–5084 (2017)
14. Shi, B., Bai, X., Belongie, S.: Detecting oriented text in natural images by linking segments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2550–2558 (2017)
15. Mery, D., Katsaggelos, A.K.: A logarithmic x-ray imaging model for baggage inspection: simulation and object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 57–65 (2017)
16. Mery, D., Svec, E., Arias, M., Riffo, V., Saavedra, J.M., Banerjee, S.: Modern computer vision techniques for x-ray testing in baggage inspection. IEEE Trans. Syst. Man Cybern. Syst. **47**(4), 682–692 (2016)
17. Uroukov, I., Speller, R.: A preliminary approach to intelligent x-ray imaging for baggage inspection at airports. Signal Process. Res. **4**, 1–11 (2015)
18. Mery, D., Riffo, V., Zscherpel, U., et al.: GDXray: The database of X-ray images for nondestructive testing. J. Nondestr. Eval. **34**(4), 42 (2015)

19. Jain, D.K.: An evaluation of deep learning based object detection strategies for threat object detection in baggage security imagery. Pattern Recognit. Lett. **120**, 112–119 (2019)
20. Cui, Y., Oztan, B.: Automated firearms detection in cargo x-ray images using RetinaNet. In: Anomaly Detection and Imaging with X-Rays (ADIX) IV, vol. 10999, pp. 105–115. SPIE (2019)
21. Bhowmik, N., Wang, Q., Gaus, Y.F.A., et al.: The good, the bad and the ugly: evaluating convolutional neural networks for prohibited item detection using real and synthetically composited X-ray imagery. arXiv preprint arXiv:1909.11508 (2019)
22. Gaus, Y.F.A., Bhowmik, N., Akcay, S., et al.: Evaluating the transferability and adversarial discrimination of convolutional neural networks for threat object detection and classification within x-ray security imagery. In: 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 420–425. IEEE (2019)
23. Wei, Y., Liu, X.: Dangerous goods detection based on transfer learning in X-ray images. Neural Comput. Appl. **32**, 8711–8724 (2020)
24. Cubuk, E.D., Zoph, B., Shlens, J., et al.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702–703 (2020)
25. Miao, C., Xie, L., Wan, F., et al.: 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, pp. 2119–2128 (2019)
26. Guo, S., et al.: Improved u-net for guidewire tip segmentation in x-ray fluoroscopy images. In: Proceedings of the 2019 3rd International Conference on Advances in Image Processing, pp. 55–59 (2019)
27. Chaudhary, A., Hazra, A., Chaudhary, P.: Diagnosis of chest diseases in x-ray images using deep convolutional neural network. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–6. IEEE (2019)
28. Huang, S., Wang, X., Chen, Y., Jie, X., Tang, T., Baozhong, M.: Modeling and quantitative analysis of xray transmission and backscatter imaging aimed at security inspection. Opt. Express **27**(2), 337–349 (2019)
29. Akcay, S., Breckon, T.P.: An evaluation of region based object detection strategies within x-ray baggage security imagery. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 1337–1341. IEEE (2017)
30. Akcay, S., Kundegorski, M.E., Willcocks, C.G., Breckon, T.P.: Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. IEEE Trans. Inf. Forensics Secur. **13**(9), 2203–2215 (2018)
31. Xie, S., Girshick, R., Dollár, P., et al.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
32. Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3085–3094 (2019)
33. Qin, X., Zhang, Z., Huang, C., et al.: Basnet: boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7479–7489 (2019)
34. Tian, Z., Chu, X., Wang, X., Wei, X., Shen, C.:FCOS: fully convolutional one-stage object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9627–9636 (2019)
35. Li, C., Li, L., Jiang, H., et al.: YOLOv6: a single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 (2022)

36. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7464–7475 (2023)