X-Ray Prohibited Item Detection with Deep Learning Architectures

Andy Huang (z5311916), Yuanyuan Zhang (z5526304), Boyang He (z5575322), Yizhu Zhou (z5602082), Yuxuan Wang (z5427969)

***Abstract*—The detection of prohibited items in X-ray security screening is crucial to prevent potential threats and ensure public safety in spaces such as airports, borders and secure facilities. However, challenges arise with increasing number of passengers, limitations with human fatigue and situations where objects are heavily overlapped with each other. In this project, we train deep learning models Faster RCNN, RetinaNet and state-of-the-art models EfficientDet and YOLOv8 to classify and detect five variations of cutters in three levels of occlusion. We demonstrate that EfficientDet and YOLOv8 significantly outperform older models Faster RCNN, RetinaNet and results from older papers in metrics like mAP. We further this by performing GradCAM analysis on layers of EfficientDet to understand how the network focuses on different regions of the image during object detection. An analysis demonstrates increased focus on sharp changes in gradient and areas where inorganic materials are present.**

Keywords—object detection, security inspection; X-ray images; occlusion

I. INTRODUCTION

A. Problem Definition

This research focuses on developing a fast and accurate neural network to assist security inspectors in detecting prohibited items in X-ray scanned images. The system aims to achieve high accuracy, robustness to varying appearances as well as the ability to distinguish between different prohibited items and backgrounds. Furthermore, it should provide correct class predictions and appropriate bounding boxes for classification and detection. In this project, we train deep architectures: Faster RCNN, RetinaNet as well as contribute to the field more state-of-the-art models like EfficientDet and YOLOv8. We test each model's efficacy and extend this by incorporating GradCAM visualizations to try to understand model predictions.

B. Significance and Applications

The significance of this project addressed the following real-world challenges in security screening:

1) ***Enhanced Public Safety***: The detection of prohibited items is essential in spaces such as airports, borders and secure facilities to prevent dangerous items from entering secure area.

2) ***Reduced Human Error****:* Security inspectors often face fatigue during long shifts. An automated system ensures consistent performance during time on duty and reduces chance of missed dangerous items.

3) ***Increased Throughput****:* As the number of passengers increases, the assistance of an automated detector reduces wait times while maintaining security standards.

4) ***Adaptability****:* This system can be deployed and made more accessible across a variety of security sensitive environments, extending the typical use in airports to train stations, government buildings or critical infrastructure facilities.

II. RELATED WORK

A. Object Detection Models for Prohibited Item Detection

This project is an object detection task and therefore requires models that can produce classification and corresponding bounding boxes. Object detection models can be categorized into several families based on their architectural approach:

1) ***Two-Stage Detector:*** Two-stage object detectors function by first genreating region proposals then classifying each proposal with a CNN. R-CNN (2014) first introduced this concept by applying high-capacity CNNs to region proposals, improving mAP by 30% compared to previous methods [1]. Fast R-CNN (2015) improved on this by extracting convolutional features from the entire image and then using an RoI pooling layer to calssify proposals in a single pass, making detection over 200 times faster with greater accuracy [2]. Faster R-CNN furthered this by introducing a Region Proposal Network (RPN) that shares features with the detection network to generate proposals, thereby eliminating the need for an extenal proposal step [3].

2) ***One-stage Detectors****:* One stage object detectors perform detection by directly predicting bounding boxes and class probabilities with a single network pass, resulting in faster inference speeds comapred to two-stage methods. YOLO (You Only Look Once) represents a family of deep architectures and introduced this ideas by dividing an image into a grid and simultaneously predicting bounding boxes and class probabilities for each cell [4]. SSD (Single Shot MultiBox Detector) improved upon YOLO by using multipel features maps at different resolutions which enabled more accurate object detection at various scales and outperforming YOLO on benchmarks like PASCAL VOL [5]. RetinaNet then addressed a key limitation of one-stage detection – class imablance, by introducing focal loss which weighed down on easy negatives and focuses the model on harder examples [6]. This signicantly improved accuracy while still maintining model efficiency.

3) ***Feature Pyramic Networks:*** Feature Pyramic Networks (FPN) were introduced to enhance object detection by creating multi-scale feature representations without significantly increasing computation cost. Traditional CNNs generated feature maps at multiple scales but only used the final layers for detection. This wrosened a networks ability to detect small objects. FPN addressed this by building a top-down architecture with alteral connections allowing detection of both higher level and lower level features. This idea also substantially improved improved performance with detectors such as Faster R-CNN [7].

4) ***More Optimised Architectures:*** More state-of-the-art developments have focused on improving both accuracy and efficiency through various architectural innovations. EfficientDET introduced a scalable and lightweight model by

combining an EfficientNet backbone with a bi-directional feature pyramid network (BiFPN), which achieved state-of-the-art accuracy whilst signifiantly reducing the number of parameters [8]. Moreover, the YOLO family had continued to improve through innovations like advanced training stratgeied, model scaling, architectural optimisation and refined pilines, with YOLOv8 being relased in 2023 by Ultralytics [9].

B. Model Explainability Techniques

As deep models became more complex, understanding their decision-making becomes essential, especially for applications involving public safety. Gradient-weighted Class Activation Mapping (Grad-CAM) is a widely used technique for visualizing and understanding deep CNNs by highlighting important regions of an image that contribute most to a model's prediction [10]. This method is particularly useful for understanding model behavior such as pinpointing class characteristics and diagnosing errors.

C. Occluded Prohibited Item X-ray Dataset

The task of detecting prohibited item in X-ray security images had been particularly challenging due to frequent object occlusions and complex visual overlaps. To address this, Wei et Al contributed the Occluded Prohibited Items X-ray (OPIXray) dataset, a high-quality benchmark designed for this task. Furthermore, they had evaluated three models SSD, YOLOv3 and FCOS on this dataset, achieving the following results [11].

Table 1: Performance comparison between SSD, YOLOv3 and FCOS on OPIXray Dataset.

Model	SSD	YOLOv3	FCOS
mAP	70.89	78.12	82.02

Our project further implements new object detection models and compares results with this table.

III. METHODS

The field of object detection has advanced rapidly with the development of new architectures that build upon traditional approaches. For our task, we develop and train a Faster R-CNN model and use it as baseline to compare with RetinaNet, EfficientDet and YOLOv8. All models use a pretrained CNN backbone and is finetuned on the dataset.

A. Faster R-CNN

In this project, we employed a Faster-RCNN with a ResNet-50 backbone. he ResNet-50 backbone serves as a feature extractor, capturing rich hierarchical representations necessary for accurate object detection, particularly in challenging scenarios with occlusions, which are common in X-ray security imagery. This served as a solid baseline due to its proven high accuracy and robust performance across various tasks.

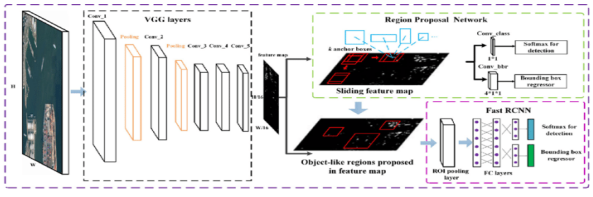

Fig. 1. Architecture of Faster R-CNN

B. RetinaNet

We also implement RetinaNet using a ResNet-50 backbone. RetinaNet has with a Feature Pyramid Network (FPN) for multi-scale feature extraction and is a one-stage detector that directly predicts object classes and bounding boxes without a separate region proposal step. Its key innovation is the introduction of focal loss, which addresses the extreme class imbalance between background and foreground objects by down-weighting easy examples and focusing the model's learning on misclassified examples.

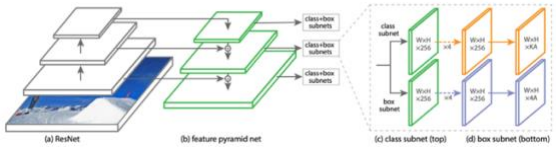

Fig. 2. Architecture of RetinaNet

C. EfficientDet with EfficientNetB0 backbone

EfficientDet leverages the highly scalable and efficient architecture, namely EfficientNet which consists of 8 CNN's designed to be computationally efficient while achieving high accuracy. It uses a concept called compound scaling which allows for balanced scaling of image size, depth and width with sizes ranging from b0 being the simplest and b7 being the largest. Our model uses a pretrained EfficientNetb0 as the backbone as it is fast to train, doesn't require much memory and still yields solid performance. It consists of three main components: an EfficientNet backbone, of which we use the simplest b0, a Bi-FPN for multi-scale feature fusion and a shared class and box prediction network as shown in **Fig.3**.

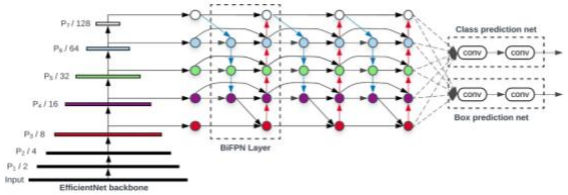

Fig. 3. Architecture of EfficientDet

Its key innovation of the BiFPN allows the network to interpret both low-level and high-level by performing a top-down pass and bottom-up pass. This makes it particularly useful for our project as X-ray images contain small cutter items within a larger luggage environment.

D. YOLOv8

Finally, we use YOLOv8n (nano), a lightweight pre-trained deep learning model which has been optimized for real-time object detection and prediction. YOLOv8n is a fully convolutional neural network (CNN) structured into three main components (see **Fig. 4**): the backbone, which extracts multi-level semantic features from the input image; the neck, which fuses features across different scales to improve the detection of both large and small objects; and the head, which predicts bounding boxes and class labels from the fused feature maps.

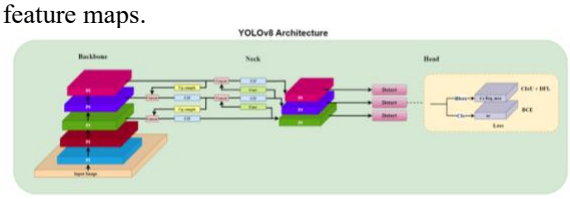

Fig. 4. Architecture of YOLOv8

IV. EXPERIMENTS

A. Dataset

This project used the OPIXray dataset consistening of 8885 X-ray images which have been scanned by security inspection machines and annotated by professional inspectors. This has been pre-split into 7019 (80%) training and 1776 (20%) testing samples. OPIXray consists of 5 categories of prohibited cutters; folding knife, straight knife, scissors, utility knife and multi-tool knife as shown in **Fig. 5.**

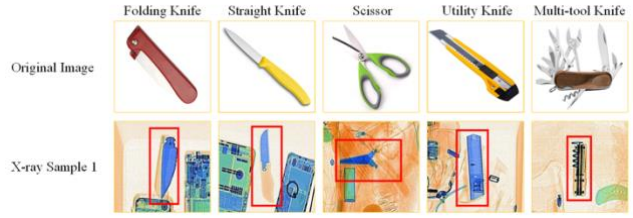

Fig. 5. Examples of Prohibited Items for Different Knife Categories

1) *Occlusion levels:*

The testing dataset has also been divided into three occlusion levels. Occlusion level 1 contains items which have no or slight occlusion, level 2 contains partial occlusion of cutters and level three contains severe or full occlusion of cutters. Visualisation can be seen in **Fig. 6**.

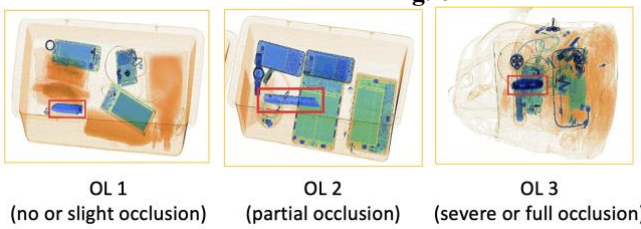

Fig. 6. illustration of Occlusion Levels (OL)

2) *Data Exploration:*

Imbalanced Data

The dataset contains significant data imbalance (**Fig. 7.**) specifically in the straight knife category which contains around half the sample as other classes, which could lead to poorer predictive performance towards this minority class if model training optimized on accuracy.

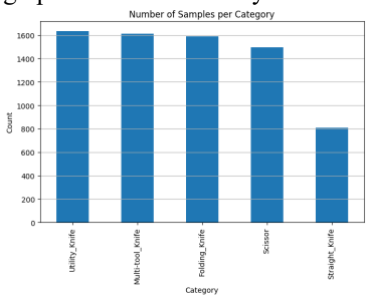

Fig. 7. Number of Samples per Category

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

Fig. 8. Number of Samples in OPIXray dataset

3) *Features of prohibited Item X-ray Images*

Shape Preservation: A core property of X-ray images is that they preserve shape properties even under occlusion. When objects overlap or are partially hodden, the edge information remains distinctly visible in the resulting image due to the penetrative mnatrues of the Xray detection machines.

Material Awareness: X-ray security systems render different materials in characteristics color schemes- inorganics materials such as metals appear in blue hues, organic materials including plastics fabrics and biological substance display orange tones while mixed materials or those with intermediate visual cues appear more green.

While these properties could be extracted and used to enhance model learning, we choose not to use this as pre-trained CNNs on datasets such as ImageNet would not have learned these features and it is uncertain whether it would lead to worse performance. Moreover, different X-ray machines may produce different color mappings and so, to produce a more generalized model, the domain-specific feature of material awareness property is not extracted.

B. Evaluation Metrics

Classification performance was evaluated on precision, recall, and F1-score. For object detection, we use Intersection over Union (IoU) which measures the overlap between predicted and ground truth bounding boxes, providing a spatial accuracy metric for localization and Mean Average Precision (mAP) which calculates the mean of Average Precision values across different object classes and IoU thresholds, combining both localization and classification performance. MAP will be used as the primary evaluation metric used across OPIXray papers.

C. Experimental Setup

To ensure experimental consistency, we standardized all input images to 512×512 resolution and employed the Adam optimizer with a constant learning rate of 0.0001 throughout training. Each model underwent identical training durations of exactly 10 epochs, and importantly, we deliberately avoided applying any data augmentation techniques in accordance with the previous paper's specifications. This was done to facilitate direct comparison of models and results with the existing papers.

V. RESULTS

Occlusion Level	YOLOv8	EfficientDet	Faster R-CNN	RetinaNet
OL1	0.8671	0.8813	0.8670	**0.8840**
OL2	0.8525	0.8649	0.8108	**0.8940**
OL3	0.8525	**0.8964**	0.7770	0.8913
Overall	0.8666	0.8790	0.8351	**0.8883**

Fig. 9. Precision Across Occlusion Levels

	Our Results				OPIXray Paper Results		
Occlusion Level	YOLOv8	EfficientDet	Faster R-CNN	RetinaNet	SSD	YOLOv3	FCOS
OL1	0.8853	0.8813	0.7034	0.7822	75.45	–	–
OL2	0.8690	0.8649	0.6820	0.7692	69.54	–	–
OL3	0.8112	0.8064	0.6481	0.7477	66.30	–	–
Overall	0.8468	0.8410	0.6861	0.7705	0.7089	0.7812	0.8202

Fig. 10. mAP Across Occlusion Levels

A. Comparison of Models

In this study, multiple object detection models, including YOLOv8, EfficientDet, RetinaNet, and Faster R-CNN, were evaluated based on their precision scores across all object categories. As illustrated in **Fig. 9**, all models demonstrated comparable precision, with RetinaNet achieving the highest overall precision among the evaluated models.

Notably, Faster R-CNN failed to detect any instances of the Straight Knife category. This outcome is likely attributed to the class imbalance present in the dataset, where the number of Straight Knife samples is significantly lower compared to other categories, as shown in **Fig. 7**. Additionally, Faster R-CNN exhibited difficulties in handling occlusion scenarios, which further impacted its detection capability for smaller or partially hidden objects. Thus, while most models maintained stable performance with high precision, Faster R-CNN was more susceptible to challenges posed by occlusion and imbalanced data distribution.

B. mAP Comparison Under Varying Occlusion Levels

The mean Average Precision (mAP) scores of the models were further analyzed under different levels of occlusion to assess robustness. As presented in **Fig. 10**, both YOLOv8 and EfficientDet achieved the highest mAP scores across all occlusion conditions, outperforming not only the other models in this study but also surpassing benchmarks reported in their original publications, where models such as SSD and YOLOv3 were evaluated.

In contrast, Faster R-CNN and RetinaNet, despite exhibiting competitive precision scores, recorded relatively lower mAP values. This indicates that while these models can precisely classify detected objects, their ability to locate and detect objects accurately, particularly under occlusion, is limited compared to YOLOv8 and EfficientDet.

These results highlight the superior adaptability of YOLOv8 and EfficientDet in complex environments involving partial object visibility.

C. Attention Visualisation Analysis

In this section, we perform GradCAM analysis on multiple layers of the trained EfficientDet architecture as seen in figure x. This revealed distinct roles across different network components. **Fig. 11** and **Fig. 12** demonstrated some backbone layers would filter potential regions containing prohibited items across broader luggage areas while Figure C and figure D would focus on individual object classification.

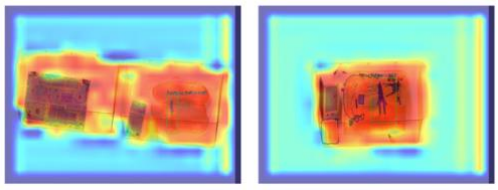

Fig. 11. Grad-CAM Visualization Result 1

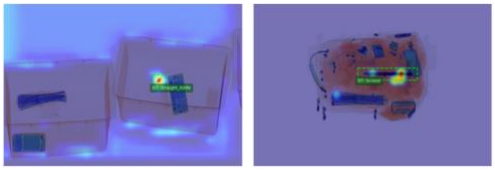

Fig. 12. Grad-CAM Visualization Result 2

While identifying unique class characteristics proved challenging due the all items being relatively small, we observed EfficientDet would yield a strong response to sharp changes in gradients as well as increased focus on blue-coloured regions throughout various layers, aligning with the material properties of metallic cutters.

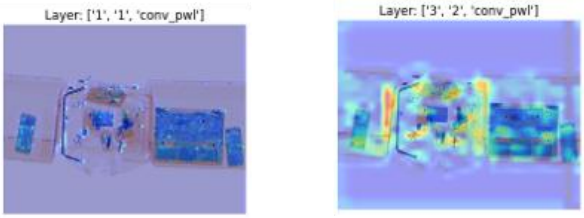

Fig. 13. Grad-CAM Visualization Result 3

These findings suggest color-specific filters or attention mechanisms could be used to better leverage the domain-specific properties of X-ray images.

VI. CONCLUSION

In this project, we conducted a comprehensive evaluation of four object detection models—YOLOv8, EfficientDet, RetinaNet, and Faster R-CNN—on detecting prohibited items in X-ray imagery. We systematically benchmarked these models under varying levels of occlusion, analyzed their performance against class imbalance, and performed attention visualizations to interpret model behavior. Our key contribution lies in identifying and leveraging models, such as YOLOv8 and EfficientDet, that demonstrate superior performance, displaying their feasibility in real-life applications. In addition, through attention visualization, we revealed domain-specific patterns, such as the tendency of models to focus on metallic regions, suggesting potential avenues for further domain adaptation strategies.

VII. IMPROVEMENTS AND FURTHER WORK

A. Data Augmentation and Imbalanced Data Techniques

The results revealed that Faster R-CNN failed to detect any instances of the straight knife category, likely due to significant class imbalance in the dataset. This limitation can be addressed through data augmentation techniques, specifically reflections and translations. Moreover, oversampling would generally make the models more robust by balancing class representation during training, thereby leading to overall improved detection performance.

B. Attention Mechanisms on Domain specific features

Incorporating attention mechanisms could significantly improve detection performance by allowing models to focus more relevant features in X-ray images. Convolutional Block Attention Module (CBAM) or Squeeze-and-Excitation (SE) blocks could be integrated into existing architectures to increase focus on more important regions. Our Grad-CAM analysis also revealed that models concentrate more on blue regions, suggesting that color-specific filters could be used to extract these domain-specific features before feeding them through attention modules and models. Additionally, transformer-based architectures like DETR on X-Ray detection may also be further explored with their built-in self-attention mechanisms.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (references)

[2] Girshick, R., 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[3] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

[4] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

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

[6] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2999-3007, doi: 10.1109/ICCV.2017.324.

[7] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 936-944, doi: 10.1109/CVPR.2017.106.

[8] M. Tan, R. Pang and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 10778-10787, doi: 10.1109/CVPR42600.2020.01079.

[9] Yaseen, M., 2024. What is yolov9: An in-depth exploration of the internal features of the next-generation object detector. arXiv preprint arXiv:2409.07813.

[10] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. International journal of computer vision, 128, pp.336-359.

[11] Wei, Y., Tao, R., Wu, Z., Ma, Y., Zhang, L. and Liu, X., 2020, October. 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).