Semi-supervised Object-Wise Anomaly Detection for Firearm and Firearm Component Detection in X-ray Security Imagery

Yona Falinie A. Gaus[1], Brian K.S. Isaac-Medina[1],
Neelanjan Bhowmik[1], Yee T. Lam[2], Toby P. Breckon[1,2]
Department of {Computer Science[1], Engineering[2]}, Durham University, Durham, UK

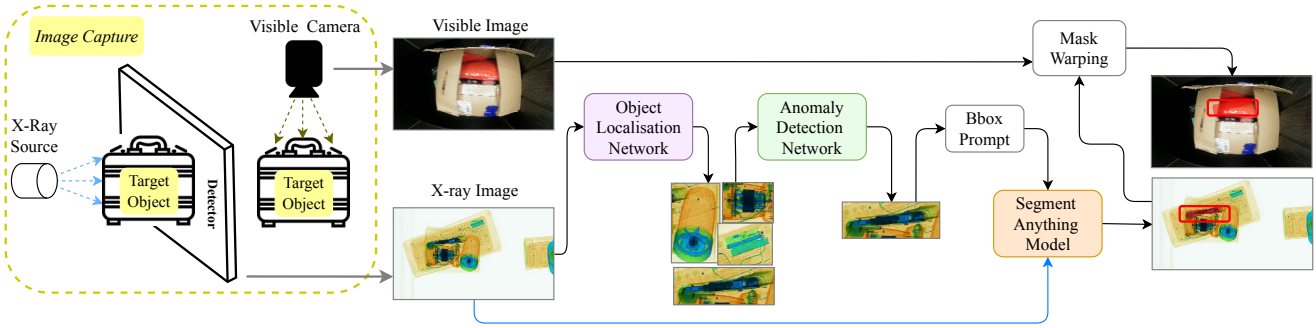


Figure 1. Our postal mail screening architecture uses dual-energy X-ray and visible-band imagery. X-ray images are processed via an Open-World Object Detector to identify potential objects, which are then processed via an Anomaly Detection Network from which detected anomalous objects are segmented using a Foundational Segmentation Model to obtain precise shape. Subsequently, detected masks are overlaid onto the corresponding visible-band image, via a cross-modal homography, for final presentation to the screening operator.

Abstract

Automatic detection of prohibited items in X-ray imagery plays a vital role in ensuring public safety, particularly in high-throughput venue, transport and postal (border) security. Existing Automatic Prohibited Item Detection Systems (APIDS), based on supervised object detection approaches, are primarily designed for venue and transport screening operations security, where individual baggage items are screened in a controlled manner. However, postal mail screening presents unique challenges due to both of the continuous flow of items, and the desire for high-throughput screening of postal mail items in an unordered and unstructured manner on existing conveyer systems, making the adaptation of current APIDS solutions impractical. To address these challenges, we propose a framework leveraging open-world object detection and semi-supervised anomaly detection as a conduit to effective screening in this context. Our approach jointly uses an open-world object detector to detect generic objects within the cluttered X-ray imagery, followed by a secondary anomaly detection network that identifies outlier objects in a class-agnostic manner. Specifically considering the context of postal screening, experimental results on a UK government evaluation dataset and a locally collected in-house Postal Mail (Parcel) dataset demonstrate the efficacy of our method, achieving high recall (77.76%) and accuracy (75.93%) with low false positive rates (1.98%), thus illustrating future potential in automated postal screening for firearms and firearms components.

1. Introduction

Firearm-related crime, including the illegal import of firearms via postal mail, poses a significant threat to UK society [8, 9, 41]. According to the most recent public report by the UK Chief Inspector of Borders and Immigration [29], UK Border Force operations at major international postal hubs recorded 1,250 firearm seizures (excluding ammunition) over the 2015/16 period. While significant progress has been made in detecting firearms and firearm component within aviation security via Automatic Prohibited Item Detection Systems (APIDS) [18, 40, 53], postal screening at scale poses a number of additional challenges in terms of volume and complexity, and hence remains reliant on an intelligence-driven manual search and seizure strategy, lacking APIDS solutions capable of screening the large variety of items in postal transit without incurring a significant false alarm rate.

APIDS is widely employed across aviation and transportation sectors to detect prohibited items by analyzing X-ray imagery of baggage, freight, and transportation hubs, via the use of X-ray security screening [1, 6, 39]. The advancement of deep learning algorithms within Convolutional Neural Network (CNN) based methods has proven to be effective in detecting a wide range of object classes

in X-ray security screening [6, 10, 11, 18, 40, 53]. In this context, several benchmarks for aviation security inspection have been developed [38, 40, 54, 58]. Whilst the current performance of APIDS heavily depends on the availability of suitable X-ray imagery datasets with sufficient object annotations, diversity and scale, public datasets such as GDXray [38], SIXray [40], PIDray [58] or OPIXray [54] have significantly advanced the development of prohibited item detection in X-ray images using CNN based methods [6, 13, 22, 36, 46, 53].

Notwithstanding the advancements in aviation security, current APIDS solutions may not translate well to a postal screening setting. While X-ray technology is utilized for postal screening at border security [48, 49], its operational deployment differs significantly from aviation security. For instance, in an aviation security environment, baggage items are screened individually in a controlled, one-by-one process at security checkpoints. By contrast, high-throughput mail screening must accommodate unstructured, mixed arrays of postal items continuously moving through mail handling facilities without incurring a significant false alarm rate. The unconstrained nature of high-throughput postal screening also gives rise to a related challenge of identifying the corresponding postal item within the mixed array of mail on the conveyor belt to the one within which a potential was detected (i.e. *"which parcel is it?"*).

As an exemplar, we specifically aim to address the challenges of firearm and firearm component detection within cluttered and complex X-ray security imagery, representative of 'stream of commerce' fast postal mail screening operations as a semi-supervised anomaly detection problem. This is achieved by leveraging two primary deep learning architectures: 1) class agnostic object detection to individually separate postal mail items on the conveyor, 2) and an anomaly detection approach to detect outlier items on an object-wise basis. Here we examine the application of the object localisation network (OLN) architecture [33], that leverages an open-world object detector (OWOD), in order to localize unseen objects without prior class supervision. Whilst classic object detectors [12, 43, 51] are focused on detecting objects from a set of known categories, OWOD [23, 33, 59] naturally capture both known and unknown objects, and hence are well suited to the unconstrained nature of objects that occur in 'stream of commerce' postal screening. A key advantage of the proposed approach is that it has no inherent requirement for labelled training data and will readily process the (noisy and cluttered) postal mail X-ray security into individual objects/regions irregardless of the nature or distribution of the highly varied items present. These individual objects/regions are subsequently passed to the secondary anomaly detection network which serves as the basis for training contemporary semi-supervised anomaly detection approaches in an object-wise manner, such that anomalous objects can be detected as *"never seen before"* in postal mail screening.

In addition, a cross-modal homography mapping between the X-ray image, within which the anomalous object has been detected, and the corresponding visible-band image of the postal mail (parcel) on the conveyance system facilitates overlay of segmented items (via [34]) onto the corresponding visible-band imagery for presentation to a security operator to assist on-conveyor identification.

2. Literature Review

APIDS are increasingly being deployed as an operational capability within aviation security, benefiting significantly from advancements in deep learning [2]. Earlier approaches use CNN either as feature extractors or classifiers, usually using small regions obtained via a sliding window approach [4, 32]. For instance, Jaccard *et al*. [32] train a CNN classifier via threat image synthesis of firearms in empty containers for cargo verification whilst [4] address a range of six prohibited items using a patch-based CNN approach. Rogers *et al*. [44] use a dual-energy CNN for firearm detection using synthetic data in a sliding window paradigm. Since the complex nature of X-ray security data makes it difficult to collect large datasets that enable training deep CNN, Akçay *et al*. [6] study the transfer learning paradigm to train fully CNN for prohibited item detection.

Several studies have used modern supervised object detection architectures for threat item detection. Akçay and Breckon [3] conducted a comprehensive evaluation of region-based detection architectures [15, 20, 21, 43]. Franzel *et al*. [17] propose a multi-view region CNN (R-CNN) to leverage the multi-view nature of X-ray security scanners in airports. Similarly, Isaac-Medina *et al*. [30] use the eipolar geometry of multiple views to constrain object detection. While these works are based on private or proprietary datasets, the availability of large-scale public datasets with extensive annotated X-ray imagery has accelerated security screening applications. For instance, SIXray [40] addresses the challenges of class imbalance and X-ray image complexities through a class-balanced hierarchical refinement approach. OPIXray [54] enhances feature representation by introducing a de-occlusion attention module to mitigate occlusions in X-ray imagery. Meanwhile, PIXray [37] contributes a diverse dataset of prohibited items and introduces a dense de-overlap attention snake for improved segmentation. These datasets have significantly advanced the development of prohibited item detection in X-ray imagery [6, 13, 22, 25–27, 36, 46, 53]. Isaac-Medina *et al*. [31] provide a comprehensive study of how these datasets are different from visible imagery datasets and how object detection architectures are affected.

Despite significant advancements in APIDS, existing methods are predominantly designed for aviation security,

with limited research addressing X-ray security screening in other transportation hubs, such as cargo and freight terminals. Notably, postal mail screening operations remain largely unexplored, leaving a critical gap in automated threat detection beyond airport security in terms of addressing the additional challenges of complexity, variety and mixed unconstrained screening presentation. Another concern is that in real-world security screening, prohibited items range from visually distinct objects to highly concealed threats, making detection inherently complex. By contrast, existing APIDS approaches assume a small set of prohibited items relying on supervised learning with known object categories to identify occurrences. This contrasts sharply with operational reality, where previously unseen (*unlabeled*) objects frequently appear, and genuine anomalies are rare, yet often important to detect, occurrences.

To address these limitations, this paper presents a framework for anomaly detection in postal mail, specifically fast parcel (UK terminology), screening operations, without dependence on predefined class labels within the 'stream of commerce' postal items being screened. Our proposed approach integrates two complementary sub-architectures: (1) Class-Agnostic Open-World Object Detection (Sec. 3.1), which identifies objects without reliance on known categories, and (2) Anomaly Detection (Sec. 3.2), which discerns deviations from normal postal mail contents.

3. Methodology

Our proposed solution consists of the combination of two architectures for different tasks that jointly enable class-agnostic object-based anomaly detection. First, an OWOD is used to detect all possible objects within the scene. Subsequently, the detected object regions are passed to the anomaly detection network to learn the normal distribution. During inference, the class-agnostic detector isolates the individual objects within X-ray imagery such that the anomaly detector identifies them as normal or abnormal.

The overall operational architecture for class-agnostic anomaly detection within the context of X-ray based postal mail screening is shown in Fig. 1. Initially, postal mail is scanned with a dual-energy X-ray scanner to capture pseudo-colour X-ray images. These images are then processed by a class-agnostic OWOD that detects all potential objects within the images (Sec. 3.1). Identified objects are subsequently isolated and passed on to an anomaly detection model (Secs. 3.2 and 3.2.1), which determines whether they are anomalous or benign. The bounding box coordinates of detected anomalies prompt a Foundational Segmentation Model [34] to generate precise object instance masks that detail the precise shape and position outline [19] of potential anomalies within each postal mail (parcel) item. Additionally, based on a wide-lens visible-band (RGB) camera at the exit of the X-ray scanner tunnel, we map these instance masks as overlays onto the visible-band images via a cross-modal homography transform, enabling the identification of the abnormal objects within the postal mail as they appear on the conveyance system from the X-ray scanner (Sec. 3.3).

3.1. Class Agnostic Open World Object Detector

In order to detect anomalies within X-ray security imagery from postal mail (parcel) screening, an object detector capable of localising all objects, including those belonging to unknown classes that are not explicitly present in any training dataset, is required. To this end, we adopt the object localisation network (OLN) architecture [33] to identify and localize objects within the image. As shown in Fig. 2, the OLN consists of a two-stage detector with a region proposal network (RPN) [43], a bounding box regression branch and an optional mask branch. In this context, given postal mail (parcel) X-ray imagery, $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, a 2D feature map $\mathbf{f} = \psi(\mathbf{x}), \mathbf{f} \in \mathbb{R}^{H' \times W' \times C}$ is extracted by a backbone network ψ. Subsequently, the RPN predicts a set of N proposal bounding boxes $P = \{p_1, p_2, \ldots, p_N\}, p_i \in \mathbb{R}^4$. Each object candidate $p_i = (x, y, w, h)$ is parameterised by one point (x, y), usually the top-left corner, and the bounding box width w and height h. Finally, the RPN in OLN regresses the centerness [47] c_i of the bounding box aiming to achieve a maximal overlap with the ground truth bounding box as a measurement of proposal quality. Additionally, the bounding box parameters of valid proposals, for example, those with an intersection over union (IoU) greater than a threshold, are also regressed. L1 losses are used for both the centerness and the box parameters. The RPN loss is subsequently thus defined as:

$$\mathcal{L}_{RPN} = \frac{1}{N} \sum_{i}^{N} \mathrm{L}_{\mathrm{L1}}(c_i, \hat{c}_i) + \mathbb{1}_{obj} \mathrm{L}_{\mathrm{L1}}(p_i, \hat{p}_i)\,, \quad (1)$$

where $\mathbb{1}_{obj}$ is 1 if the proposal is matched against a ground truth (0 otherwise). Subsequently, the proposal features $\mathbf{u}_i$ are extracted from $\mathbf{f}$ using RoIAlign [43]. These features are then fed into a shared network $g()$ producing an object representation $\mathbf{v}_i = g(\mathbf{u}_i) \in \mathbb{R}^d$. Similarly to the RPN, a bounding box branch regresses the box quality b_i and the bounding box parameters p using L1 losses. The bounding box loss is then:

$$\mathcal{L}_{bbox} = \frac{1}{N_v} \sum_{i}^{N_v} \mathrm{L}_{\mathrm{L1}}(b_i, \hat{b}_i) + \mathrm{L}_{\mathrm{L1}}(p_i, \hat{p}_i)\,, \quad (2)$$

where N_v is the number of valid proposals and the box quality b_i is measured by the IoU with the ground truth.

The bounding box confidence is calculated as $s_i = \sqrt{(c_i b_i)}$, and these bounding boxes, which constitute benign objects, $\mathcal{D}$ will form an input set to our proposed anomaly detection model in the next section.

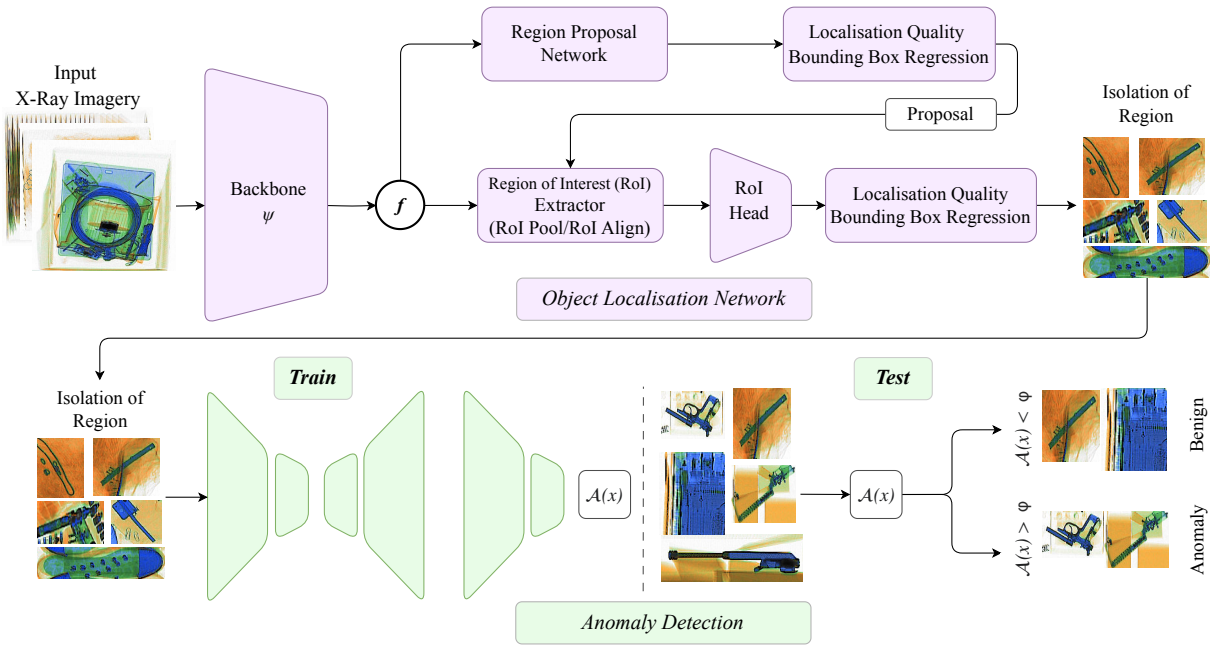


Figure 2. Our proposed architecture for open-world anomaly detection combines the object localisation network (OLN), and semi-supervised anomaly detection network.

3.2. Anomaly Detection

Given the relatively large volumes of non-anomalous (normal) objects returned from the previous stage, the initial goal is to model this distribution of benign (object) image samples, $\mathcal{D}$, in order to learn a feature embedding (i.e., manifold), such that we can subsequently detect abnormal samples, $\hat{\mathcal{D}}$, as outliers during the inference stage.

Each anomaly detection model, $f()$, learns the data distribution of benign (object) image samples by minimising the anomaly score $\mathcal{A}(x)$ over the training data set, itself comprising benign object instances only. For a given test image $\hat{x}$, a high anomaly score of $\mathcal{A}(\hat{x})$ indicates the presence of anomalies within the image. The evaluation criteria for this is to threshold (φ) the score, where $\mathcal{A}(\hat{x}) > \varphi$ indicates the presence of an anomaly.

3.2.1. Semi-Supervised Anomaly Detection Approaches

A contemporary paradigm to addressing anomaly detection in numerous types of imagery is to perform algorithm training based solely on normal data samples (i.e. images of normality) in what is often denoted as a *"semi-supervised"* approach [5]. Given the challenges of comprehensive anomalous data collection within the context of postal screening, here we leverage this approach by down-selecting approaches that do not require anomalous exemplars for training. In particular, we explore four families of approaches: (i) one-class feature-based classification, (ii) reconstruction-based distribution learning, (iii) flow-based model and (iv) student-teacher based methodology; that form the basis for the down-selected set of anomaly detection algorithms subsequently evaluated for the task in hand.

One-class classification [7] refers to methods that learn a feature manifold for normal data while constraining the manifold to be as compact as possible. At test time, any data mapped outside the learned manifold is classified as anomalous. Deep Feature Kernal Density Estimation (DFKDE) [7] is a one-class anomaly classification algorithm that includes a deep learning-based feature extraction using a pre-trained ResNet-50 [28] backbone and an anomaly classification stage featuring Principal Component Analysis (PCA) and Gaussian Kernel Density Estimation (KDE). Initially, normal images are processed through ResNet-50 to produce a feature vector of length 2048 from the average pooling layer. These features are then reduced to 16 principal components via PCA, which transforms high-dimensional data into a lower-dimensional space retaining most variance. In the final stage, Gaussian KDE models the distribution of these PCA-reduced features. During inference, anomalies are detected when the probability density falls below a preset threshold, indicating the presence of an anomaly against the data distribution learned from the training dataset.

Reconstruction-based approaches [5] [56] [57] use autoencoder neural network architectures in order to learn how to reconstruct images from the normal data distribution via a compressed encoder-decoder representation. These are similarly trained exclusively on normal data samples and such that a poor reconstruction of a given example at test time is used to detect the presence of anomalous image regions. Generative Adversarial Network Anomaly (GANomaly) [5] utilises a GAN to successfully recreate normal regions while failing to handle anomalies. These methods are trained exclusively on images without anoma-

lies and typically involve manual post-processing steps to pinpoint the anomalies, which limits the potential to optimise feature extraction for enhanced detection efficiency. Conversely, Discriminatively Trained Reconstruction Anomaly Embedding Model (DRAEM) [56] and Dual Subspace Re-projection Network (DSR) [57] learn a combined representation of an anomalous image and its normal reconstruction, while also establishing a clear decision boundary between normal and anomalous cases. This method allows for direct localisation of anomalies without requiring complex post-processing, and it can be trained with straightforward and broad-based anomaly simulations.

Flow-based approaches [55] [45] are used to learn transformations between data distributions. This is achieved by first extracting features that do not contain anomalies from a pre-trained network and mapped by a trainable flow model to fit a uniform Gaussian distribution. In Unsupervised Anomaly Detection and Localisation via 2D Normalising Flows (FastFlow) [55] a two-dimensional normalising flow is employed to independently process feature maps at each scale and averages the results at the end. Fully Convolutional Cross-Scale Normalizing Flow (CS-Flow) [45] introduces a new kind of normalising flow model that jointly processes multiple feature maps across different scales. For each of the flow-based models, during the testing phase, the normalising flow model is employed to accurately assess the likelihood of a test image. Images that are anomalous will typically fall outside this distribution and therefore exhibit a lower likelihood compared to normal images.

In the student-teacher methodology [16] [52], the teacher acts as the feature extractor during the training phase and imparts knowledge to the student model. Reverse Distillation (RD) [16] and Student Teacher Feature Pyramid Network (STFPM) [52] use a single pair of teacher encoder and student decoder networks. In RD, the student network does not directly process raw images; instead, it uses the class embedding from the teacher model as input and aims to reconstruct the teacher multi-scale representations. For both models, during inference, the teacher generates features that are unfamiliar to the student if an abnormal image is presented, hindering the student network from accurately replicating these features. Therefore, the disparity in features between the teacher and the student networks becomes the key factor for identifying anomalies.

3.3. Cross-modal Homography Recovery

In parallel to X-ray postal mail scanning, visible images of the same postal mail (parcel) items are captured using a wide-lens visible-band (RGB) camera fitted at the exit of the X-ray scanner tunnel. The camera is placed such that it is parallel to the imaging plane of the X-ray scanning system. As both the X-ray and visible images represent the same objects lying in the same plane (the X-ray scanner belt plane), we employ a homography transform from the X-ray imaging plane to that of the visible camera, such that the predicted instance masks from the X-ray images can be readily mapped onto the visible images in the corresponding position of occurrence. We formulate this transformation from the X-ray imaging plane to that of the camera as:

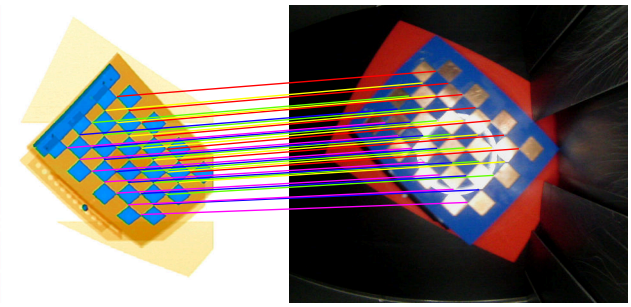


Figure 3. Cross-modal feature point correspondences are obtained across a number of exemplar images in both the X-ray and visible-band imaging modalities using a tailored cross-modal calibration target, from which homography estimation is then performed.

$$x_{visible} = \mathbf{H} x_{xray}\,, \tag{3}$$

where x is the point in homogeneous coordinates from either of the X-ray or imaging source and $\mathbf{H}$ is a 3×3 homography matrix [24]. The homography transform, $\mathbf{H}$ between the two imaging planes is recovered via least square error minimization of the feature reprojection error from matched features across imagery from both domains based on the use of a specially designed cross-modal calibration target, as shown in Fig. 3.

Ultimately, the homography matrix is used to map the position of potential anomalies detected in the X-ray image to the corresponding position in the visible image, as illustrated in Fig. 1. This integration ensures that anomalies detected in X-ray images are accurately highlighted in the corresponding visible images, enhancing operator ability to identify and assess these irregularities.

4. Results

This section presents the dataset used for evaluation, implementation details and final evaluation results.

4.1. Datasets

We perform our quantitative evaluation using UK government evaluation dataset (available upon request from UK Home Office / UK Defence Science Technology Laboratory (DSTL)) [50]. This dataset comprises of both expertly concealed firearm (threat) items and operational benign (non-threat) imagery from commercial X-ray postal mail security screening operations on the UK. with a total of 207,337 training images (benign) and 4,500 test images (threat). The firearm (threat) item comprise of three categories, full weapon (firearm in its normal, operational form), set (full weapon disassembled in to its base components) and indi-

vidual (single component). For our experiments, we consider only the individual category, which consists of firearm components {*barrel, bolt assembly, bolt carrier assembly, casing, central block, cocking handle, magazine, shotgun internals, slide, spring, spring pin, trigger assembly*}.

The dataset encompasses images depicting single objects as well as complex scenarios with multiple objects, providing diverse and challenging samples for training and evaluation. Subsequently, we perform our qualitative evaluation using a locally constructed in-house postal mail dataset, comprising a set of parcels with a subset containing disassembled firearms as {*firearm components*}.

4.2. Implementation Details

Within the UK government evaluation dataset [50], we select the benign (*i.e.* non firearm) item as the training set, whilst the firearm threat containing items are used for testing. Since our anomaly detection framework only accepts object-based bounding boxes, we first extract a large number of objects by training OLN [33] (Sec. 3.1) pre-trained on MS-COCO dataset [35] and fine tuned on SIXRay10 [40], via the MMDetection framework [14], and therein only retain the localisation information and discard the classification labels, following [33]. The OLN is trained via Stochastic Gradient Descent with a learning rate of 0.00252, momentum of 0.9 and weight decay of 10^{-4}. Finally, we train the network for 100 epochs with a batch size as 16.

We construct object-centric training examples from detected objects in the benign imagery, according to predicted bounding boxes. The final dataset consists of a total of 240,000 benign object items, which are used for training. To explore the impact of the training set size, and to enhance the utility of the experiment, the training dataset is further divided into two subsets: a small set with 80,000 images and a full set containing all 240,000 images. We rescale all images to a fixed size (256×256) and adopt Anomalib [7] as the base framework for anomaly detection. The test set consists of 4,500 images with an unannotated threat items present. In this sense, if our pipeline finds at least one anomalous objects during inference, the whole bag is labelled as an anomaly instance. Therefore, our evaluation metrics are based on normal/anomalous bags based on this rule, since no ground truth annotations are available.

All implementations and visualisation are conducted in PyTorch [42] framework with a single NVIDIA 1080Ti GPU. We use the Segment Anything Model (SAM) [34] for building instance masks, using the detected bounding box as prompts. For fair comparison and consistency, we use the same parameters for all experiments; the parameters follow the defaults used in [7] or within the original work.

4.3. Evaluation Results

Tabs. 1 and 2 report the recall, false positive rate (FPR), accuracy and area under the receiving operating characteristic (AUROC) curve over the UK government evaluation dataset [50], as explained in Sec. 3.2. For AUROC calculations, and since we use object detections to label bags as normal/abnormal (Sec. 4.2), we consider two versions with different anomaly score strategies: AUROC_M, where the bag takes the maximum anomalous score detected, and AUROC_A, where the bag takes the average anomaly score. We use a detection confidence threshold of 0.6 for the OLN. The reported FPR score uses an anomaly score threshold such that 95% of the normal instances are tagged as normal. In this context, a positive example refers to normal.

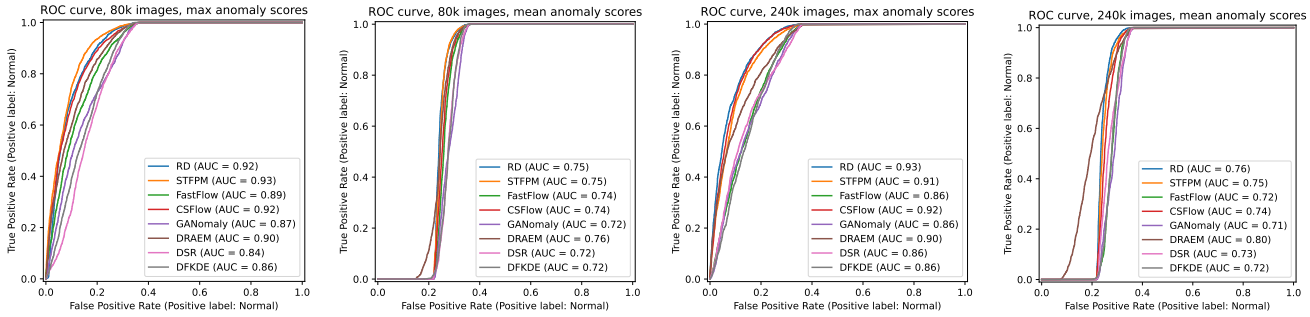
Method	Model	↑ Recall (%)	↓ FPR (%)	↑ Accuracy (%)	↑ AUROC_M	↑ AUROC_A
Student-Teacher	RD [16]	45.56	21.11	62.85	0.916	0.754
	STFPM [52]	65.22	14.14	**75.93**	**0.929**	0.752
Flow	FastFlow [55]	60.42	16.16	72.57	0.889	0.737
	CSFlow [45]	36.60	**3.53**	67.66	0.915	0.743
Reconstruction	GANomaly [5]	**77.76**	57.08	59.68	0.866	0.719
	DRAEM [56]	66.62	31.70	67.49	0.903	**0.756**
	DSR [57]	65.02	20.55	72.51	0.841	0.722
One-Class	DFKDE [7]	58.09	18.78	70.09	0.858	0.722

Table 1. Performance of anomaly detection using a smaller set of training images (80k) tested on the firearm components test set.

Method	Model	↑ Recall (%)	↓ FPR (%)	↑ Accuracy (%)	↑ AUROC_M	↑ AUROC_A
Student-Teacher	RD [16]	28.76	2.21	64.57	**0.927**	0.757
	STFPM [52]	50.62	8.08	72.05	0.913	0.751
Flow	FastFlow [55]	39.33	11.30	64.94	0.863	0.721
	CSFlow [45]	22.91	**1.98**	61.88	0.921	0.743
Reconstruction	GANomaly [5]	69.47	21.38	**74.22**	0.858	0.714
	DRAEM [56]	59.44	23.62	68.23	0.896	**0.796**
	DSR [57]	**76.33**	56.48	59.31	0.864	0.727
One-Class	DFKDE [7]	43.11	13.46	65.64	0.857	0.721

Table 2. Performance of anomaly detection using a larger set of training images (240k) tested on the firearm components test set.

Tab. 1 shows the anomaly detection using the small set of benign training images (80,000 images) and tested on firearm component images (4,500 images). In detail, CS-Flow gives the lowest FPR of 3.53% but only achieves an accuracy of 67.66% due to a small recall being of 36.60%. FastFlow and DFKDE also give reasonable FPR as low as 16.16% and 18.78% and higher accuracy at 72.57% and 72.51%, respectively. STFPM achieves a balanced performance, delivering a recall of 65.22% and an FPR of 14.14%, achieving 75.93% of accuracy, which is further supported by having the highest AUROC_M of 0.929. Whilst there is a considerable difference in performance for most models, it is observed that reconstruction methods, such as GANomaly, DRAEM and DSR, usually achieve high recall, but high FPR. This may be attributed to some anomaly instances being significantly small in size (*spring, spring pin*) and the models are not able to successfully differentiate these small-scale anomalies from general reconstruction noise when applied to complex and cluttered X-ray security imagery. The key of STFPM success could be attributed the hierarchical feature matching strategy, enabling the student network to receive a mixture of multi-scale knowledge from the feature pyramid under stronger supervision from



(a) (b) (c) (d)

Figure 4. ROC curves of different methods varying the predicted anomaly score. (a) 80k images partition, max anomaly score, (b) 80k images partition, avg anomaly score, (c) 240k images partition, max anomaly score,(d) 240k images partition, avg anomaly score,.

the teacher and thus more successfully enables the detection of anomalies at varying scales. Figs. 4a and 4b shows the ROC curves for maximum and average anomaly scores. A consistently improved performance for reconstruction-based methods is observed for the maximum ROC curves. On the other hand, averaged anomaly scores shows a sharp transition in the ROC curve. Since anomalous bags still have normal items, averaging the anomaly scores would remove the impact of a single detected anomaly, which is why the AUROC_A score shows a decreased performance.

Similarly, Tab. 2 shows the anomaly detection performance on full sets of benign training images (240,000 images). Except from DSR, a considerable decrease of the FPR is observed for all methods when using the full 240k dataset, although the recall is also decreased. This effect is observed because having a more diverse set of normal objects might cause more confusion between normal and abnormal objects. Similar to the 80k objects case, reconstruction models achieve high recall but high FPR. Notably, student-teacher models achieve high AUROC_M, with 0.927 and 0.913 for RD and STFPM, with low FPRs of 2.21% and 8.08%, respectively. Flow-based approaches such as CSFlow also reported lower FPR of 1.98%, overall 61.88% accuracy and AUROC_M of 0.921. Additionally, GANomaly has an increased performance with an accuracy of 74.22%, with a reasonable FPR, highlighting that this method is benefitted from increased training data. The different behaviour of each model underscores the trade-offs inherent in these methods. Finally, Figs. 4c and 4d shows the ROC curves for this training set partition. While RD, STFPM and CSFlow are the best models as with the 80k instances case (Fig. 4a), the separation is clearer, highlighting these methods as their effectiveness within our pipeline. A similar trend to the 80k the averaged ROC curve is observed.

Qualitative results are illustrated in Fig. 5 using a locally constructed in-house postal mail (parcel) dataset that uses X-ray security imagery collected via a Gilardoni (FEP ME 640 AMX) dual-energy x-ray scanner and comprising various firearm and firearm components as representative anomalies within an otherwise benign 'stream of commerce' set of parcel examples. We can also see that we have effectively visualized the spatial position of the detected anomalous items on the visible-band image of the parcels as they appear on the conveyance system using the homography-based geometric mapping from the X-ray imagery and corresponding visible-band (RGB) images. In some cases, the predicted masks can be noisy (Fig. 5, 6th row), an effect that has been previously documented for foundational models in X-ray imagery [19]. The use of better foundational models in this image modality is still an open research direction.

5. Conclusion

In this work, we introduce a semi-supervised class-agnostic anomaly detection framework tailored for postal mail (parcel) screening operations. By integrating open-world object detection and semi-supervised anomaly detection, our approach effectively identifies prohibited items without the need for extensive annotated datasets whilst also coping with the variety of complexity of object occurrence within 'stream of commerce' postal mail. The experimental results highlight the robustness of our methodology with normalizing flow based anomaly detection methods consistently achieving low false positive reporting whilst reconstruction based method are able to achieve higher recall, in detecting a wide variety of firearm/firearm components under challenging conditions. This approach not only addresses the limitations of existing APIDS for postal mail screening operations but also paves the way for further advancements in anomaly based threat item detection. Future work will focus on enhancing the scalability of the system, exploring advanced anomaly detection techniques, and expanding the framework to other security-critical domains.

Acknowledgement: This work was funded via the Defence and Security Accelerator (UK) under the Innovative Research Call in Explosives and Weapons Detection (2023), sponsored by UK National Protective Security Authority, UK Department for Transport, UK Defence Science and Technology Laboratory, UK Home Office, Metropolitan Police Service, US Department of Homeland Security, (Science & Technology Directorate, DHS), UK Border Force and UK Ministry of Justice.

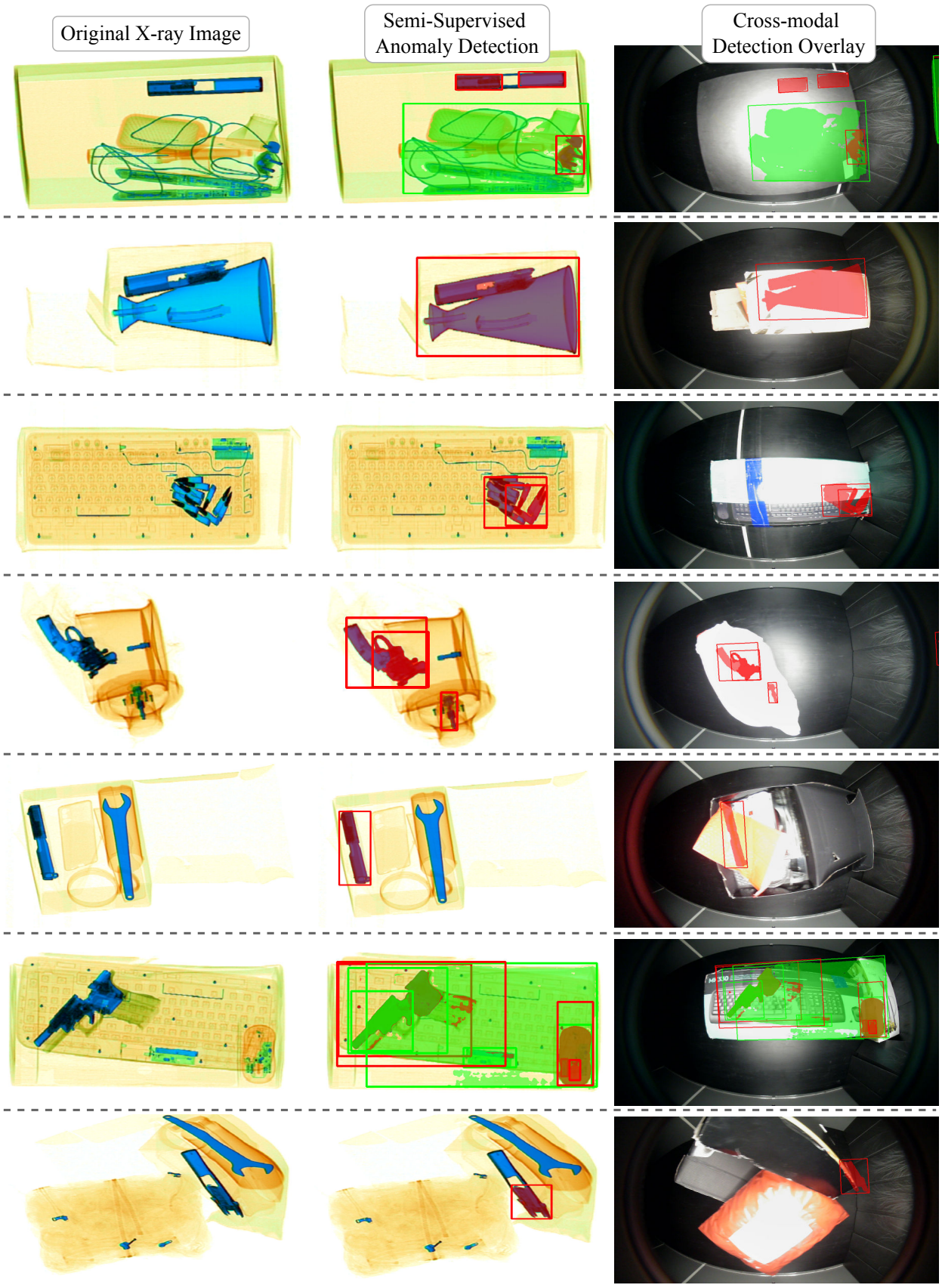


Figure 5. Exemplar anomaly detection from in-house postal mail (parcel) dataset. The first column shows X-ray images captured using a Gilardoni (FEP ME 640 AMX) dual-energy X-ray scanner. The second column visualizes the detected anomalous items (in red) and benign items (in green) overlaid on the X-ray imagery. The third column presents the corresponding visible-band (RGB) images with the detected anomalous item shape outlines mapped via a cross-modal homography transform.

References

[1] Abdelfatah Ahmed, Divya Velayudhan, Taimur Hassan, Mohammed Bennamoun, Ernesto Damiani, and Naoufel Werghi. Enhancing security in x-ray baggage scans: A contour-driven learning approach for abnormality classification and instance segmentation. *Engineering Applications of Artificial Intelligence*, 130:107639, 2024. 1

[2] Samet Akcay and Toby Breckon. Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. *Pattern Recognition*, 122:108245, 2022. 2

[3] Samet Akcay and Toby P Breckon. An evaluation of region based object detection strategies within x-ray baggage security imagery. In *IEEE Int. Conf. Image Process.*, pages 1337–1341. IEEE, 2017. 2

[4] Samet Akcay, Mikolaj E. Kundegorski, Michael Devereux, and Toby P. Breckon. Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery. In *IEEE Int. Conf. Image Process.*, pages 1057 –1061. IEEE, 2016. 2

[5] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian Conf. on Computer Vision*, pages 622–637. Springer, 2018. 4, 6

[6] Samet Akcay, Mikolaj E Kundegorski, Chris G Willcocks, and Toby P Breckon. Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. *IEEE Trans. Inform. Forensics and Security*, pages 2203–2215, 2018. 1, 2

[7] Samet Akcay, Dick Ameln, Ashwin Vaidya, Barath Lakshmanan, Nilesh Ahuja, and Utku Genc. Anomalib: A deep learning library for anomaly detection. In *IEEE Int. Conf. Image Process.*, pages 1706–1710. IEEE, 2022. 4, 6

[8] BBC News. Gun crime: How do weapons appear on england's streets? https://www.bbc.co.uk/news/uk-44053904, 2018. Accessed: 2025-02-24. 1

[9] BBC News. Northern ireland: Police seize firearms in major operation. https://www.bbc.co.uk/news/uk-northern-ireland-63987375, 2022. Accessed: 2025-02-25. 1

[10] Neelanjan Bhowmik, Qian Wang, Yona F.A. Gaus, Marcin Szarek, and Toby P. Breckon. The good, the bad and the ugly: Evaluating convolutional neural networks for prohibited item detection using real and synthetically composite x-ray imagery. In *Proc. British Machine Vision Conference Workshops*, pages 1–8. BMVA, 2019. 2

[11] Neelanjan Bhowmik, Yona Falinie A Gaus, and Toby P Breckon. On the impact of using x-ray energy response imagery for object detection via convolutional neural networks. In *IEEE Int. Conf. Image Process.*, pages 1224–1228. IEEE, 2021. 2

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229. Springer, 2020. 2

[13] An Chang, Yu Zhang, Shunli Zhang, Leisheng Zhong, and Li Zhang. Detecting prohibited objects with physical size constraint from cluttered x-ray baggage images. *Knowledge-Based Systems*, 237:107916, 2022. 2

[14] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[15] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Adv. Neural Inform. Process. Syst.*, pages 379–387, 2016. 2

[16] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9737–9746, 2022. 5, 6

[17] Thorsten Franzel, Uwe Schmidt, and Stefan Roth. Object detection in multi-view x-ray images. In *Pattern Recognition*, pages 144–154, 2012. 2

[18] Yona Falinie A Gaus, Neelanjan Bhowmik, Samet Akçay, Paolo M Guillen-Garcia, Jack W Barker, and Toby P Breckon. Evaluation of a dual convolutional neural network architecture for object-wise anomaly detection in cluttered x-ray security imagery. In *Int. Joint Conf. Neural Networks*, pages 1–8. IEEE, 2019. 1, 2

[19] Yona Falinie A Gaus, Neelanjan Bhowmik, Brian KS Isaac-Medina, and Toby P Breckon. Performance evaluation of segment anything model with variational prompting for application to non-visible spectrum imagery. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 3142–3152, 2024. 3, 7

[20] Ross Girshick. Fast r-cnn. In *Int. Conf. Comput. Vis.*, pages 1440–1448, 2015. 2

[21] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, page 580–587. IEEE Computer Society, 2014. 2

[22] Bangzhong Gu, Rongjun Ge, Yang Chen, Limin Luo, and Gouenou Coatrieux. Automatic and robust object detection in x-ray baggage inspection using deep convolutional neural networks. *Trans. on Industrial Electronics*, pages 10248–10257, 2020. 2

[23] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9235–9244, 2022. 2

[24] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 5

[25] Taimur Hassan and Naoufel Werghi. Trainable structure tensors for autonomous baggage threat detection under extreme occlusion. In *Asian Conf. on Comput. Vis.*, 2020. 2

[26] Taimur Hassan, Meriem Bettayeb, Samet Akçay, Salman Khan, Mohammed Bennamoun, and Naoufel Werghi. Detecting prohibited items in x-ray images: A contour proposal learning approach. In *IEEE Int. Conf. Image Process.*, pages 2016–2020. IEEE, 2020.

[27] Taimur Hassan, Muhammad Shafay, Samet Akçay, Salman Khan, Mohammed Bennamoun, Ernesto Damiani, and Naoufel Werghi. Meta-transfer learning driven tensor-shot detector for the autonomous localization and recognition of concealed baggage threats. *Sensors*, 20(22):6450, 2020. 2

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 4

[29] Independent Chief Inspector of Borders and Immigration. An Inspection of Border Force Operations at Coventry and Langley Postal Hubs, March–July 2016, 2017. [Online; accessed 7 Feb. 2024]. 1

[30] Brian KS Isaac-Medina, Chris G Willcocks, and Toby P Breckon. Multi-view object detection using epipolar constraints within cluttered x-ray security imagery. In *2020 25th International conference on pattern recognition (ICPR)*, pages 9889–9896. IEEE, 2021. 2

[31] Brian KS Isaac-Medina, Seyma Yucer, Neelanjan Bhowmik, and Toby P Breckon. Seeing through the data: A statistical evaluation of prohibited item detection benchmark datasets for x-ray security screening. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 524–533, 2023. 2

[32] Nicolas Jaccard, Thomas W Rogers, Edward J Morton, and Lewis D Griffin. Tackling the x-ray cargo inspection challenge using machine learning. In *Anomaly Detection and Imaging with X-Rays (ADIX)*, pages 131–143. SPIE, 2016. 2

[33] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *Proc. Robotics and Automation Letters (RA-L)*, 2022. 2, 3, 6

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Int. Conf. Comput. Vis.*, pages 4015–4026, 2023. 2, 3, 6

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014. 6

[36] Zhongqiu Liu, Jianchao Li, Yuan Shu, and Dongping Zhang. Detection and recognition of security detection object based on yolo9000. In *Int. Conf. on Systems and Informatics*, pages 278–282. IEEE, 2018. 2

[37] Bowen Ma, Tong Jia, Min Su, Xiaodong Jia, Dongyue Chen, and Yichun Zhang. Automated segmentation of prohibited items in x-ray baggage images using dense de-overlap attention snake. *IEEE Trans. on Multimedia*, pages 1–1, 2022. 2

[38] Domingo Mery, Vladimir Riffo, Uwe Zscherpel, German Mondragón, Iván Lillo, Irene Zuccar, Hans Lobel, and Miguel Carrasco. Gdxray: The database of x-ray images for nondestructive testing. *J. of Nondestructive Evaluation*, 34(4):42, 2015. 2

[39] Domingo Mery, Erick Svec, Marco Arias, Vladimir Riffo, Jose M Saavedra, and Sandipan Banerjee. Modern computer vision techniques for x-ray testing in baggage inspection. *IEEE Trans. on Sys., Man., and Cybernetics: Systems*, 47 (4):682–692, 2016. 1

[40] Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao, and Qixiang Ye. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2119–2128, 2019. 1, 2, 6

[41] National Crime Agency. Firearms - national crime agency. https://www.nationalcrimeagency.gov.uk/what-we-do/crime-threats/firearms, 2025. Accessed: 2025-02-25. 1

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, pages 8024–8035, 2019. 6

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*, pages 91–99, 2015. 2, 3

[44] Thomas W Rogers, Nicolas Jaccard, and Lewis D Griffin. A deep learning framework for the automated inspection of complex dual-energy x-ray cargo imagery. In *Proc. Anomaly Detection and Imaging with X-Rays (ADIX) II*, pages 106–117. SPIE, 2017. 2

[45] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *IEEE Winter Conf. on Appli. of Comput. Vis.*, pages 1088–1097. IEEE, 2022. 5, 6

[46] Malarvizhi Subramani, Kayalvizhi Rajaduari, Siddhartha Dhar Choudhury, Anita Topkar, and Vijayakumar Ponnusamy. Evaluating one stage detector architecture of convolutional neural network for threat object detection using x-ray baggage security imaging. *Rev. d'Intelligence Artif.*, 34(4):495–500, 2020. 2

[47] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Int. Conf. Comput. Vis.*, pages 9627–9636, 2019. 3

[48] UK Home Office. Response to an Inspection of Border Force's Fast Parcels Operations. https://www.gov.uk/government/publications/response-to-an-inspection-of-border-forces-fast-parcels-operations/response-to-an-inspection-of-border-forces-fast-parcels-operations, 2023. [Online; accessed 7 Feb. 2024]. 2

[49] UK Home Office. An inspection of border force's fast parcels operations: May to july 2023. UK Government Report, 2024. Accessed: 2025-02-25. 2

[50] UK Home Office Centre for Applied Science and Technology (CAST). OSCT Borders X-ray Image Library. Online, 2016. Publication Number: 146/16. 5, 6

[51] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7464–7475, 2023. 2

[52] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for unsupervised anomaly detection. In *Brit. Mach. Vis. Conf.*, 2021. 5, 6

[53] Thomas W Webb, Neelanjan Bhowmik, Yona Falinie A Gaus, and Toby P Breckon. Operationalizing convolutional neural network architectures for prohibited object detection in x-ray imagery. In *IEEE Int. Conf. on Machine Learning and Appli.*, pages 610–615. IEEE, 2021. 1, 2

[54] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In *ACM Int. Conf. Multimedia*, page 138–146, 2020. 2

[55] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021. 5, 6

[56] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. A discriminatively trained reconstruction embedding for surface anomaly detection. In *Int. Conf. Comput. Vis.*, pages 8330–8339, 2021. 4, 5, 6

[57] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr–a dual subspace re-projection network for surface anomaly detection. In *Eur. Conf. Comput. Vis.*, pages 539–554. Springer, 2022. 4, 5, 6

[58] Libo Zhang, Lutao Jiang, Ruyi Ji, and Heng Fan. Pidray: A large-scale x-ray benchmark for real-world prohibited item detection. *Int. J. Comput. Vis.*, 131:3170–3192, 2023. 2

[59] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2