Seeing Through the Data: A Statistical Evaluation of Prohibited Item Detection Benchmark Datasets for X-ray Security Screening

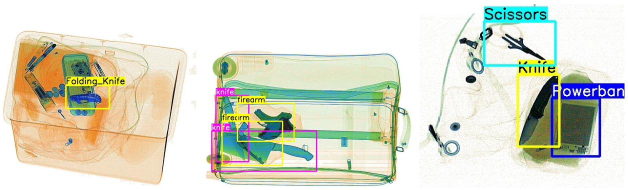
Brian K. S. Isaac-Medina[1], Seyma Yucer[1], Neelanjan Bhowmik[1], Toby P. Breckon[1, 2]
Department of {[1]Computer Science, [2]Engineering}, Durham University, UK

Abstract

The rapid progress in automatic prohibited object detection within the context of X-ray security screening, driven forward by advances in deep learning, has resulted in the first internationally-recognized, application-focused object detection performance standard (ECAC Common Testing Methodology for Automated Prohibited Item Detection Systems). However, the ever-increasing volume of detection work in this application area is highly reliant on a limited set of large-scale benchmark detection datasets that are specific to this domain. This study provides a comprehensive quantitative analysis of the underlying distribution of the prohibited item instances in three of the most prevalent X-ray security imagery benchmark and how these correlate against the detection performance of six state-of-the-art object detectors spanning multiple contemporary object detection paradigms. We focus on object size, location and aspect ratio within the image in addition to looking at global properties such as image colour distribution. Our results show a clear correlation between false negative (missed) detections and object size with the distribution of undetected items being statistically smaller in size than those typically found in the corresponding dataset as a whole. For false positive detections, the size distribution of such false alarm instances is shown to differ from the corresponding dataset test distribution in all cases. Furthermore, we observe that one-stage, anchor-free object detectors may be more vulnerable to the detection of heavily occluded or cluttered objects than other approaches whilst the detection of smaller prohibited item instances such as bullets remains more challenging than other object types.

1. Introduction

X-ray security screening is widely used in aviation and other transportation domains, with a recent focus on the development of automatic identification of prohibited items within complex and cluttered X-ray images using a range of object detection approaches [1]. These developments have now led to changes in international aviation security regulations resulting in the first international security equipment standard for automatic prohibited item detection - the European Civil Aviation Conference (ECAC) Common Testing Methodology for the integration of Automated Prohibited Item Detection Systems (APIDS), which provides certified performance compliance for X-ray security scanner systems in the area of automated threat object detection (ECAC APIDS) and possibly represents one of the first, if not the first, internationally recognised performance standard for object detection algorithm performance [33].



OPIXray *SIXray10* *PIDray*

Figure 1. Typical images from X-ray datasets SIXray [26], OPIXray [40] and PIDray [41].

Within this context, prior work has investigated the performance of deep learning-based detectors for security inspection and threat-item detection within X-ray security imagery [3, 8, 15, 22, 36, 39]. Furthermore, recent work has seen the introduction of new paradigms for object detection, such as the use of Vision Transformers [23] and anchor-free models [12, 42, 43]. However, the performance of all of these object detection approaches is very dependent on the availability of suitable X-ray security imagery datasets with sufficient object annotations, diversity and scale which has often been lacking within the common public X-ray dataset resources [1, 25, 28].

Previous works have investigated the use of transfer learning to overcome the relatively small size of X-ray security datasets for image classification [2] and object detection [11, 39] and report that a pre-trained model on a large-scale dataset such as ImageNet [32] or MS-COCO [21] results in higher detection performance despite the cross-over from perspective projection photographic imagery to parallel projection transmission imagery. However, pre-training on such datasets could induce dataset bias that may not hold for the target dataset [13] which exhibits many differences from photographic image (object detection) datasets

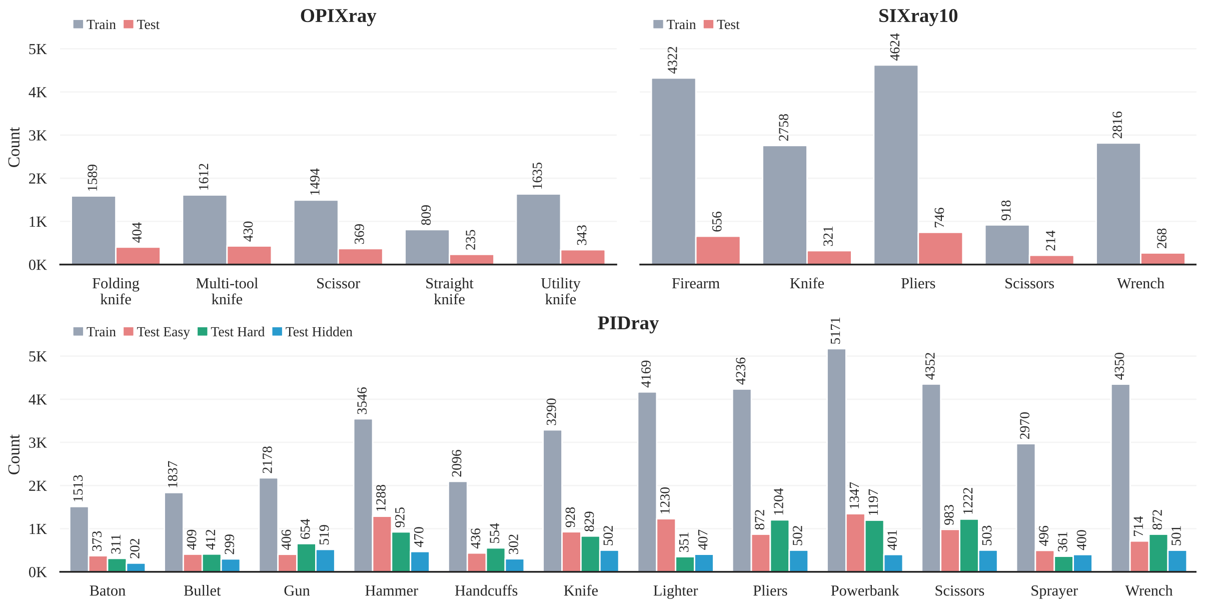


Figure 2. PIDray, OPIXray, SIXray10 dataset statistics: class-wise prohibited item instances within {*Train, Test*} data splits.

(Fig. 1). For instance, X-ray images are semi-transparent transmission imagery, meaning that objects appear translucent and visually blended front-to-back whereas, in natural photographic images, foreground objects visually occlude background objects. As a result, the creation of dedicated X-ray security datasets has been an important step in the development of APIDS-capable approaches but in itself is inherently challenging due to the requirement for concurrent access to an X-ray security scanner, a diverse range of suitable prohibited threat items and similarly a suitably diverse set of passenger bags in which to em-place them. As a result, a limited number of large-scale benchmark datasets have emerged [26, 31, 40, 41] upon which the relative performance analysis of APIDS capable approaches is now largely reliant [1, 5, 25, 28, 39]. Consequently, a statistical review of these benchmark dataset resources and their differences from more conventional object detection benchmark datasets [21], is an important step in improving the effectiveness of object detectors when applied to X-ray security prohibited item detection.

Beyond the specifics of X-ray imagery, multiple studies [16, 27, 38] provide ample evidence of dataset bias on common object recognition datasets, causing an inclination towards highly biased object detection models. In this regard, dataset bias refers to systematic errors in a dataset affecting the generalisation ability of learning-based algorithms, resulting in poor performance on models developed beyond the original dataset domain (distribution mismatch between dataset and the task) [37, 38]. The majority of the methods for object detection bias mitigation utilise dataset re-sampling to adjust the relative frequencies of dataset samples, improving the model generalisation performance [6, 19, 20]. For instance, *REPAIR* [19] removes the representation bias by learning a probability distribution over the dataset that favours hard instances for a given representation. On the other hand, *AFLITE* [18] introduces adversarial filters designed to detect different types of dataset bias to eliminate noisy labels and feature distribution skewness before training the model.

Despite the study of dataset bias becoming particularly relevant for prohibited object detection, existing studies [4, 13, 35, 46] on dataset bias have been conducted on natural photographic (visible spectrum) datasets, such as the PASCAL Visual Object Classes [10], ImageNet [32] or COCO [21]. Furthermore, as the prohibited object detection literature commonly adopts pre-trained contemporary detection architectures [2, 3], there is an increasing possibility of encountering the aforementioned dataset biases and risks in X-ray security imagery.

Against this background, in this study, we analyze the underlying statistical trends of the image samples and object instances within the most extensive, and commonplace, X-ray security imagery datasets and their resultant impact on a suite of representative object detectors, providing extensive quantitative analysis on failure modes and potential sources of detection bias.

Our key contributions are as follows:

– A statistical evaluation of three of the most extensive and commonly used X-ray security imagery benchmark datasets, namely OPIXray [40], SIXray [26] and PIDray [41], based on image and object instance properties, including image colour and object bounding box (location) distribution, highlighting the key differences against a standard natural image dataset (COCO [21]).

- A reference performance benchmark of six contemporary object detectors spanning different paradigms:- Cascade R-CNN [7] (multi-stage), Deformable DETR [45] (Transformer-based detection head), FSAF [44] (anchor-free and online feature selection), Faster R-CNN [30] with Swin Transformers [23] (two-stage detector with a Vision Transformer-based backbone), YOLOX [12] (state-of-the-art one stage real-time) and CenterNet [42] (keypoint-based).
- A quantitative investigation on the failure modes of the six different object detectors considered showing a correlation of the false negative and false positive detection occurrences against ground truth for the purpose of detection bias identification. Additionally, a class-wise analysis of the distribution of object instances within the training and testing sets for further understanding of detector performance and bias.

2. Evaluation Methodology

We present our evaluation methodology spanning down-selected datasets (Sec. 2.1), object detectors (Sec. 2.2) and object instance statistical analysis (Sec. 2.3 in addition to implementation details (Sec. 2.4).

2.1. Datasets

To assess the performance and potential dataset bias of X-ray security imagery, we analyse three of the most extensive, and commonly used, prohibited item detection datasets which are characteristically diverse, covering different X-ray scanners, prohibited item distribution and reflective of a likely real-world scenario.

OPIXray [40] consists of $8,885$ X-ray images with five classes of prohibited item (*folding knife, straight knife, scissor, utility knife, multi-tool knife*) and represents cluttered and overlapping stream-of-commerce baggage items.

SIXray [26] consists of $1,059,231$ images with $8,929$ X-ray images containing at least one prohibited item among five classes (*gun, knife, wrench, pliers, scissors*) originating from stream-of-commerce baggage and parcel X-ray scans collected from several subway stations. In this work, the *SIXray* partition is used, containing the $8,929$ images with prohibited items and $10\times$ images without.

PIDray [41] is a large-scale prohibited items dataset including 12 classes of prohibited items (*baton, bullet, gun, hammer, handcuffs, knife, lighter, pliers, power bank, scissors, sprayer, wrench*) and $124,486$ images coming from three different scenarios (airports, subway stations and railway stations). The testing partitions are divided into easy (exactly one prohibited item), hard (two or more objects in the same image) or hidden (purposely hidden objects within the bag contents).

The distribution of prohibited items object within these the datasets is illustrated in Fig. 2 with a comparison of their colour characteristics further shown in Fig. 3.

2.2. Object Detection

To provide our performance benchmark, we down-select six state-of-the-art object detection architectures spanning differing detection paradigms (e.g. single-stage, multi-stage, deep convolutional neural networks, vision Transformers).

Cascade R-CNN (CR-CNN) [7]: is a modification of the R-CNN [14] that resolves the trade-off of having to choose between low Intersection over Union (IoU) thresholds that generate imprecise detections and high IoU thresholds that negatively affect performance. It does so by training a sequence of detectors one after the other, each with a progressively higher IoU threshold, to become more discerning in identifying false positives.

FSAF [44]: is a single-stage object detection framework that uses feature selection on multiple anchor-free branches to overcome issues with heuristic-based feature selection and overlap-dependent anchor sampling. FSAF is built on a feature pyramid architecture and has been shown to improve object detection accuracy with minimal additional inference time.

Deformable DETR (DDETR) [45]: is an extension of the Detection Transformer (DETR) object detection model, which uses a transformer architecture to model sequential relationships between features that uses a deformable attention mechanism. Deformable DETR improves convergence by having attention modules focus only on adjacent features and addresses the issue of detecting objects at different scales. It retains the benefits of DETRs transformer-based architecture while achieving these improvements.

Faster R-CNN w/ Swin Transformer (FRCNNw/ST) [23]: Liu *et al.* introduced the Swin Transformer, a vision Transformer with shifted windows, which shows significant detection performance gains when used as a backbone for object detection. It is used in conjunction with Faster R-CNN [30], an anchor-based two-stage detector that uses a region proposal network.

YOLOX [12]: follows the success of the YOLO family of detectors, and is an anchor-free architectural variant of YOLOv3 [29] consisting a decoupled detection head (*i.e.*, separated networks for classification and bounding box regression) and a strong label assignment and achieves state-of-the-art performance at real-time (YOLOX-S version).

CenterNet [42]: converts the detection task to a keypoint detection by predicting the centre of the objects and regressing the remaining parameters. It achieves a great speed-accuracy trade-off and can be used for other tasks such as 3D and keypoint detection.

Table 1. Detectors training details.

Architecture	Optimiser	Epochs	Lr
CR-CNN [7]	SGD	20	10^{-2}
FSAF [44]	SGD	20	10^{-2}
DDETR [45]	Adam [17]	50	10^{-4}
FRCNNw/ST [23]	AdamW [24]	30	10^{-4}
YOLOX [12]	SGD	20	10^{-3}
CenterNet [42]	SGD	20	2×10^{-3}

2.3. Object Instance Analysis

In order to investigate the effect of the underlying distribution of object instances on detector performance, a statistical analysis of the distribution of three spatial parameters is performed: object area, centre and aspect ratio. In this context, *area* of an object refers to the total number of pixels that its bounding box occupies; *centre* is the geometrical centroid of the bounding box relative to image and *aspect ratio* is the ratio of width to height. Regarding the centre, we report the Euclidean distance from the image centre. Our analysis aims to uncover the distribution of the location and size of objects within the sample images and how this potentially differs from a natural images dataset such as COCO [21]. Furthermore, the distribution of these parameters for false positive and false negative detection results is also performed.

2.4. Implementation Details

The training of the detector architectures (Sec. 2.2) is implemented using the MMDetection framework [9]. All detectors are pre-trained on the COCO dataset [21]. Training details are implemented using the default configurations with a few modifications, shown in Tab. 1.

Standard data augmentation techniques as described in the original works are used. All training is carried out using an NVIDIA GeForce RTX 2080 Ti.

3. Evaluation Results

We present our evaluation spanning dataset analysis (Sec. 3.1), detection performance (Sec. 3.2) and detection relative to dataset object instance distributions (Sec. 3.3).

3.1. Dataset Analysis

The colour analysis of the X-ray datasets compared to the COCO dataset is shown in Fig. 3 in the form of RGB and HSV histograms. It is observed from the RGB histogram that while the COCO dataset has a seemingly uniform distribution across the intensity values, X-ray datasets are highly skewed to high values on the three RGB and HSV channels (mostly because of the white background). OPIXray and PIDray show higher peaks at 255 since they have large background regions. In contrast, SIXray10, where baggage images tend to occupy the full image plane, shows a peak at slightly smaller values, corresponding to the green, blue and orange colours of a typical bag (this peak is also observed for OPIXray and PIDray, albeit significantly lower). Additionally, the hue component distribution on the COCO dataset shows peaks at the orange (most likely corresponding to a range of lighter skin tones, since *person* is the most common category) and blue (sky in outdoor images) colours, whilst the saturation mostly decreases towards bright colours, with one peak at high saturation values, indicating a high relatively presence of pure colours. On the other hand, the X-ray datasets are generally not saturated images with peaks at the blue and orange colours, having an additional peak with a hue component of zero (corresponding to the white background).

Table 2. AP @ IoU=0.5 comparison for the *OPIXray* dataset.

Model	Folding	Straight	Scissor	Utility	M-tool	mAP
CR-CNN	0.934	0.771	0.961	0.836	0.949	0.890
FSAF	0.821	0.804	0.956	0.805	0.868	0.851
DDETR	0.909	0.774	0.963	0.859	0.934	0.888
FRCNNw/ST	0.945	0.842	0.977	0.854	0.959	0.915
YOLOX	0.908	0.801	0.974	0.859	0.935	0.896
CenterNet	0.911	0.758	0.977	0.820	0.909	0.875

Table 3. AP @ IoU=0.5 comparison for the *SIXray10* dataset.

Model	Firearm	Knife	Wrench	Pliers	Scissors	mAP
CR-CNN	0.882	0.824	0.838	0.882	0.873	0.860
FSAF	0.894	0.776	0.792	0.885	0.898	0.849
DDETR	0.913	0.934	0.910	0.944	0.960	0.932
FRCNNw/ST	0.897	0.856	0.899	0.920	0.947	0.904
YOLOX	0.909	0.869	0.891	0.907	0.938	0.903
CenterNet	0.906	0.862	0.887	0.918	0.908	0.896

The object parameters distribution is presented in Fig. 4. The dimensions, centre, aspect ration and area, are shown as contour plots, where each contour represents the probability mass of lying among different density levels (10%, 30%, 50%, 70% and 90%) with densities obtained via Gaussian kernel density estimation. It is observed from the area and dimensions distributions (Fig. 4, upper two rows) that the COCO dataset has a higher concentration of small objects, while X-ray datasets have clear peaks at 10^4 pixels. This variation is explicable in relation to the perspective image view of the COCO images that gives rise to perspective foreshortening (i.e. object further away appear smaller) whilst the parallel projection of the X-ray scan alleviates any such perspective effects. Ultimately, pre-training on the COCO dataset may leverage this prior information and hence a bias to predict small objects can be induced (see Sec. 3.3). The distribution of the object bounding box centresreveal that while objects tend to appear near the image centre in all datasets, they are constrained into the scanned region in the X-ray datasets, with OPIXray being

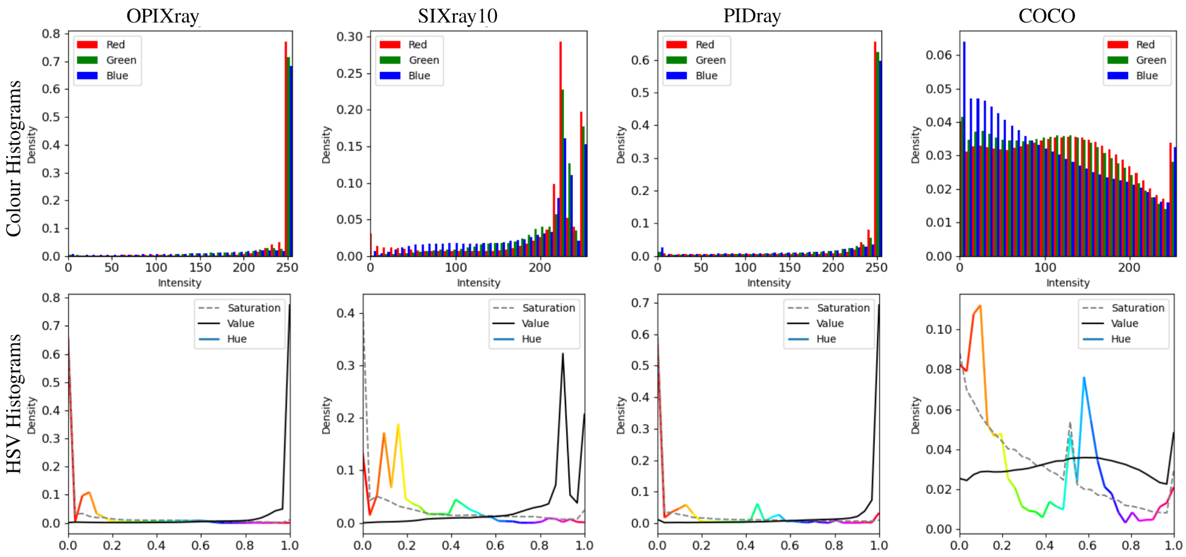


Figure 3. RGB and HSV histograms for X-Ray datasets: OPIXray [40], SIXray10 [26] and PIDray [41]; compared with COCO dataset [21].

Table 4. AP @ IoU=0.5 comparison for the *PIDray* dataset. Three reported values are evaluated on {*easy/hard/hidden*} test sets.

Model	Baton	Pliers	Hammer	Powerbank	Scissors	Wrench	Gun	Bullet	Sprayer	HandCuffs	Knife	Lighter	mAP
CR-CNN	.985/.933/.357	.999/.965/.916	.960/.898/.774	.953/.951/.753	.958/.926/.735	.984/.969/.930	.158/.416/.655	.945/.873/.332	.775/.892/.544	.989/.983/.989	.379/.630/.479	.843/.741/.125	.827/.848/.633
FSAF	.982/.940/.357	.999/.970/.890	.965/.906/.719	.952/.965/.672	.924/.931/.621	.979/.957/.942	.088/.307/.550	.950/.909/.264	.748/.866/.595	.988/.982/.990	.279/.615/.474	.855/.765/.114	.809/.843/.599
DDETR	.989/.952/.589	.999/.983/.941	.971/.945/.860	.969/.968/.723	.970/.968/.845	.987/.983/.981	.099/.337/.645	.966/.877/.384	.950/.914/.703	.988/.986/.990	.578/.724/.537	.872/.781/.388	.861/.868/.716
FRCNNw/ST	.988/.976 /.717	.990/.979/.949	.988/.952 /.921	.969/.978 /.835	.981/.963/.910	.988/.987/.990	.506/.579/.756	.962/.872/.505	.958/.943/.676	.988/.986/.990	.692/.753/.620	.867/.787/.906	.906/.896/.765
YOLOX	.986/.958/.615	.989/.986/.883	.969/.943/.826	.964/.966/.737	.982/.964/.840	.958/.987/.978	.334/.472/.666	.960/.902/.393	.905/.928/.676	.989/.986/.990	.670/.707/.525	.846/.795/.213	.879/.883/.695
CenterNet	.977/.935/.935	.990/.975/.914	.972/.908/.655	.952/.955/.649	.967/.933/.649	.983/.970/.963	.278/.441/.568	.891/.748/.207	.732/.863/.334	.989/.987/.989	.439/.605/.362	.851/.723/.143	.835/.837/.566

the most constrained case (given the small size of bags in this dataset). Additionally, the distribution of the test sets is presented. A careful examination exhibits small variances in the area between the test and training sets on the SIXray10 and PIDray datasets, while other object parameters retain similar distributions. Finally, no significant difference is found with respect to aspect ratio.

3.2. Detection Performance

The detection performance across the OPIXray, SIXray10 and PIDray datasets is shown in Tables 2 - 4. In the X-ray security detection context, being able to detect an object is more important than how accurate the bounding box is, hence we report class-wise average precision (AP) and mean AP (mAP) across all classes considering an IoU threshold of 0.5. In general, Transformer-based detectors achieve the highest detection performances, with Faster R-CNN w/Swin Transformers illustrating superior detection for the OPIXray and PIDray datasets, and Deformable DETR on SIXray10. On the other hand, FSAF and CenterNet detectors perform the weakest. On an analysis of the test splits of PIDray (Tab. 4), it is further observed that these two detectors have a significantly lower mAP for the hidden (heavily occluded object) test split, making them unreliable object detectors within this context. Interestingly, the mAP does not exhibit a notable change between the easy and hard splits (some classes increase their AP while others decrease it), indicating that the evaluated detectors are not heavily affected by the number of objects in them (the hard split contains exclusively more than one item). This is also observed by Song *et al*. [34]. Additionally, some categories are more difficult to detect than lesser dangerous objects (*e.g*., *Gun* vs *Wrench* in PIDray), demonstrating that a class-wise analysis is needed in order to create tailored object detectors that identify more important items.

3.3. Detection Performance Instance Analysis

The distributions of the ground truth bounding box properties presented in Sec. 3.1, including area, centre and aspect ratio, indicate that there is no significant distribution variance within the training and testing X-ray security datasets. Accordingly, we question *Can the detectors perform reliably on objects that belong to the same training distribution? If not, how do the predictions vary across the selected object instance parameters?* Subsequently, we evaluate the distribution of selected properties within False Negative (FN) and False Positive (FP) predictions from the chosen detectors and demonstrate the skewness of these distributions within training and testing splits (Fig. 5). Regarding the *area*, it is observed that the median value of the area of FN samples across all datasets and detectors is smaller than that of the test and train distributions, indicating that undetected objects tend to have a smaller area (pixels) compared to the ground truth set area. In addition, the distribution of

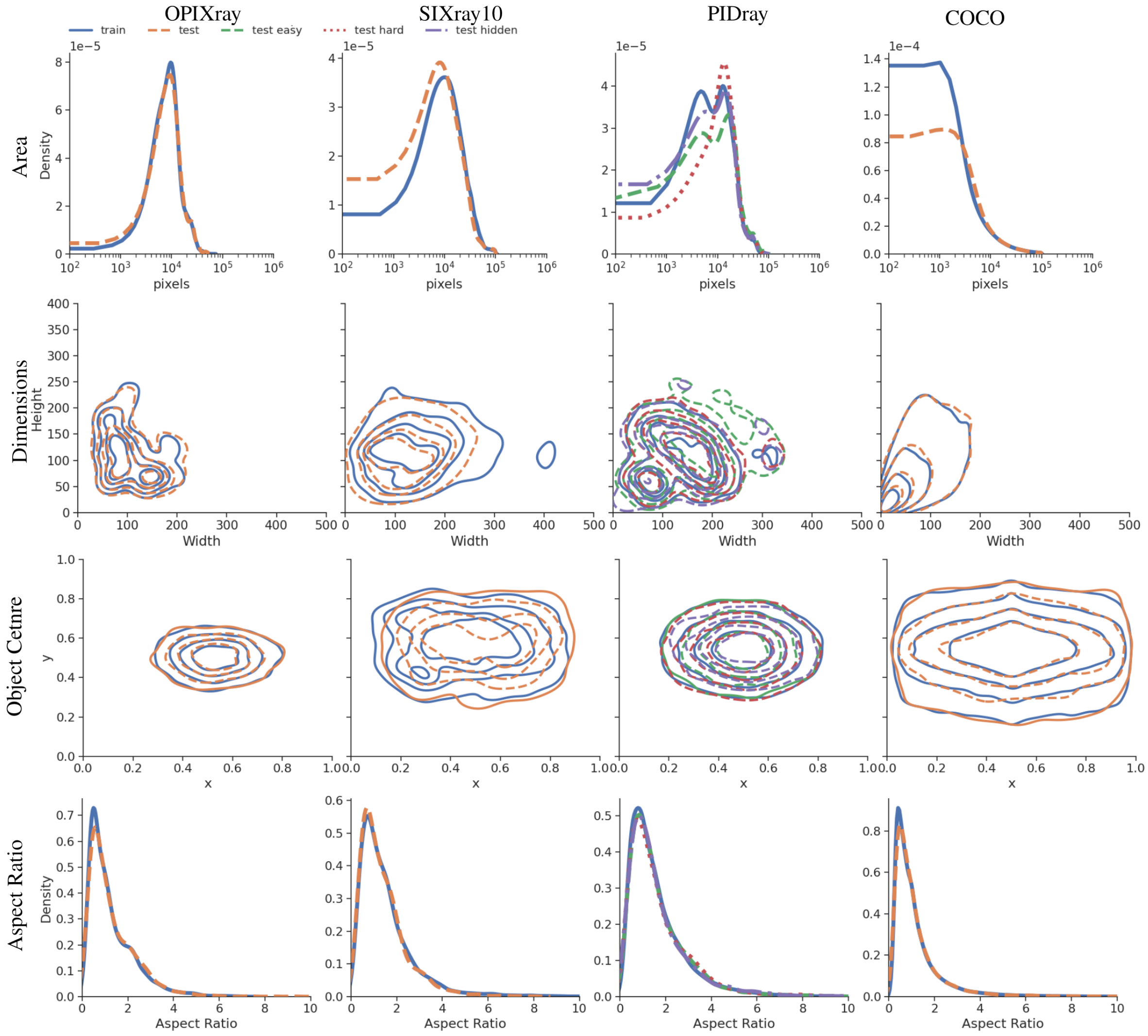


Figure 4. Density estimation (using a Gaussian kernel) of the area, dimensions, bounding box centre and aspect ratio of the ground truth bounding boxes on OPIXray, SIXray10, PIDray and COCO.

area in FP samples differs from the test distribution, with lower or higher variations depending on the detector and datasets. Notably, the FSAF detector on the PIDray Hidden (heavily occluded) set shows the most significant difference, where higher area size samples are mismatched. Conversely, the smallest distribution difference between the test and training sets was observed in the OPIXray dataset, resulting in smaller changes in predictions regarding their area. Concerning the *centre* parameter, we observed a slight increase in the median value of the distance of the FP predictions centre location from the centre of the image on the OPIXray dataset, while the rest did not exhibit any obvious trend. This indicates that while objects are usually constrained within an enclosed region, this does not affect modern detectors. As for the *aspect ratio*, the FN distribution in the OPIXray dataset shows a larger spread in aspect ratios than in the test set.

Furthermore, we explore the distribution shifts towards properties within class-level object bounding boxes within the datasets. As the area distribution exhibits the most significant changes in predictions, we focus our investigation on this parameter via the use of the PIDray test set (since it is the most challenging). Specifically, we first calculate the median area values of each class in the train, test, FN, and FP prediction sets. Subsequently, the relative error of the median $(1 - (median_{set}(FP)/median_{area}(test)))$ of FN, FP and train ground truth with respect to the test ground truth is calculated (Tab. 5), enabling us to determine *the relative change of the area among object categories regarding the evaluated sets*. Accordingly, negative values indicate that larger areas were miss-matched (FP/Test), or undetected (FN/Test), while positive values refer to smaller

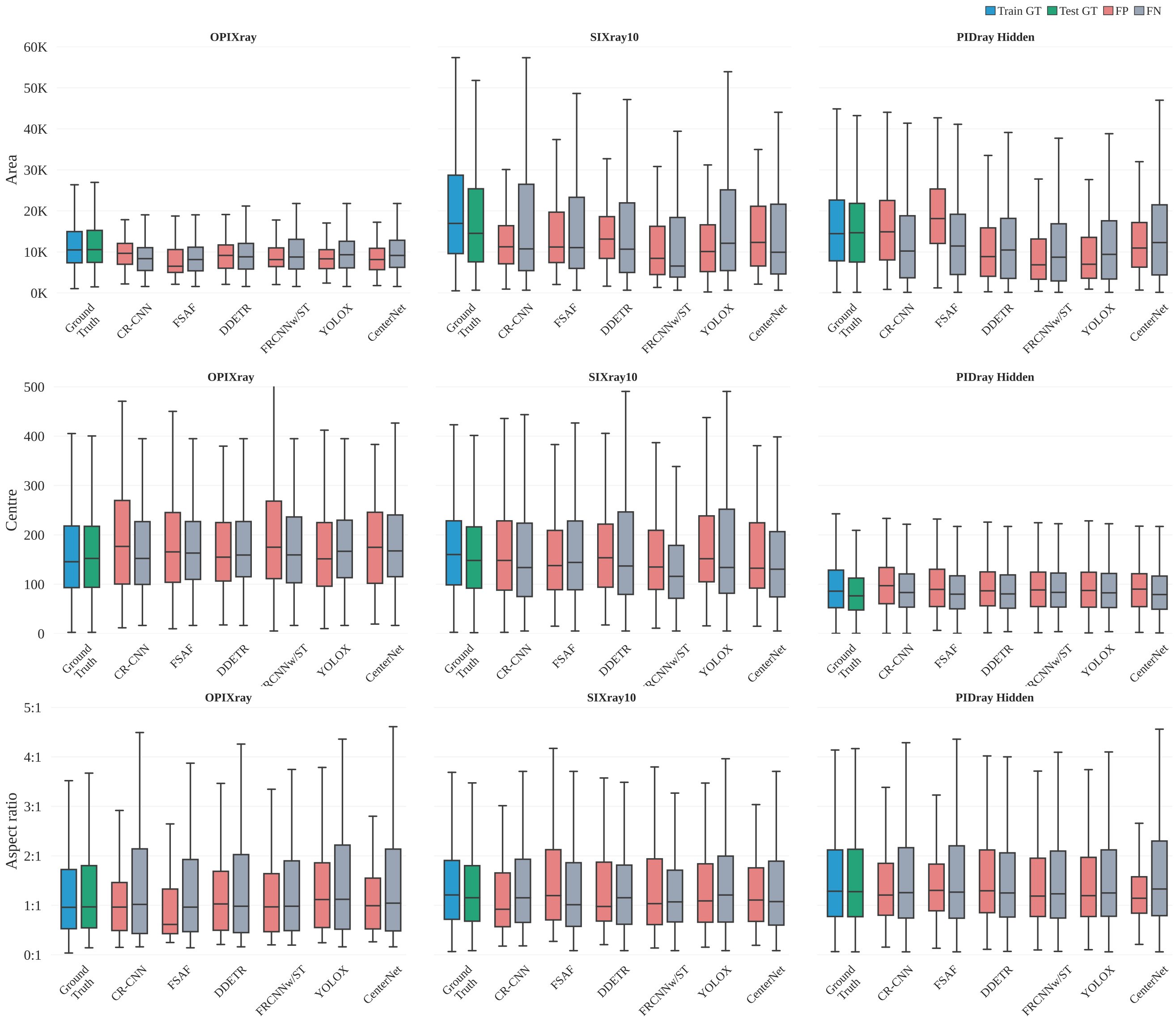


Figure 5. The distribution of detector performance across object instance parameters regarding the evaluated sets including train, test ground truth, predicted false positive and false negative sets.

area predictions compared to test distribution within these classes.

From Tab. 5, it is seen that the FP predictions for the bullet object category tend to be mismatched with larger area bounding boxes in all three PIDRay test sets. This can be explained given that bullets have small ground truth bounding boxes and small variations in the predicted bounding boxes give rise to high IoU. Conversely, wrenches are mismatched against smaller objects in the PIDray hidden (heavily occluded) data spit. It should be noted however that as some classes have fewer FN and FP depending on their performance, as the wrench category (Tab. 4). With respect to the gun category, the distribution of the FN in the hidden set is significantly smaller, meaning that either the detector cannot locate highly cluttered guns and/or that they are just partially detected with smaller bounding boxes, having a similar problem with the IoU as in the bullets (but not as drastic). Finally, the highest difference is found in the FN for Faster R-CNN w/Swin Transformer on the handcuff category of the PIDray hidden set. This, however, corresponds to a single instance and is attributable to handcuffs being the only deformable object (due to the linking chain between the bracelets), resulting in variable object geometry and hence bounding box annotations.

Table 5. The Area Percentile Change on categories of PIDray {*hidden, hard, easy*} sets from top to bottom, each cell depicts the $1 - (median(set)/median(test))$ meaning that red colour cells have larger change of the area among object categories.

Detector	Set	Baton	Bullet	Gun	Hammer	HandCuffs	Knife	Lighter	Pliers	Powerbank	Scissors	Sprayer	Wrench
GT	Train	-0.09	-0.41	0.07	0.19	0.08	0.02	-0.16	-0.24	0.15	-0.12	0.12	0.05
CR-	FN	0.17	0.54	-0.81	-0.02	0.40	0.13	0.03	0.51	0.15	0.19	-0.03	0.34
CNN	FP	0.04	-0.62	-0.12	0.23	0.62	0.44	-0.21	-0.39	-0.26	0.02	0.27	0.85
FSAF	FN	0.09	0.26	-0.02	-0.05	0.20	0.13	0.02	0.51	0.18	0.07	-0.01	0.15
	FP	0.27		-0.04	0.22	0.65	0.52	-0.45	-0.38	-0.35	0.29	0.22	0.92
DDETR	FN	0.18	0.59	-0.87	-0.11	0.23	0.21	0.09	0.52	0.07	0.21	-0.01	0.45
	FP	-0.06	0.24	0.33	0.34		0.67	-0.01	0.06	0.14	0.33	0.23	0.80
FRCNNw/ST	FN	0.18	0.68	-1.16	-0.08	-24.26	0.29	0.07	0.24	0.04	0.28	0.02	0.86
	FP	0.35	-0.40	0.56	0.51	0.49	0.61	0.03	0.06	0.40	0.44	0.32	0.67
YOLOX	FN	0.21	0.58	-0.95	-0.16	0.01	0.27	0.03	0.38	0.10	0.24	-0.08	0.65
	FP	0.40	-0.03	0.30	0.40	0.78	0.46	0.10	0.19	0.29	0.44	0.40	0.72
CenterNet	FN	0.09	0.34	-0.62	-0.08	0.09	0.13	0.03	0.41	0.10	0.11	-0.01	0.42
	FP	0.41	-0.11	0.25	0.44	-0.20	0.42	0.02	0.07	0.30	0.15	0.30	0.58
GT	Train	0.02	-0.01	0.56	0.22	-0.15	-0.17	-0.10	-0.13	0.00	0.16	0.06	0.04
CR-	FN	0.13	0.42	-0.05	0.13	0.47	0.01	0.19	0.09	0.09	0.28	-0.03	0.27
CNN	FP	0.51	-2.43	0.59	0.48	0.00	0.04	-0.09	-0.05	0.10	0.08	0.22	0.11
FSAF	FN	0.19	0.25	-0.04	0.01	0.29	0.04	0.18	0.14	0.22	0.19	0.02	0.25
	FP	0.15	-2.22	0.71	0.35	-0.11	-0.10	-0.06	0.08	0.02	0.14	0.40	0.12
DDETR	FN	0.08	0.35	-0.05	0.06	0.25	-0.05	0.14	0.01	0.03	0.26	0.05	0.25
	FP	0.30	-1.54	0.50	0.40	-0.04	0.42	0.18	0.05	0.21	0.33	0.25	0.05
FRCNNw/ST	FN	0.13	0.48	-0.05	0.08	0.33	-0.09	0.16	0.09	0.11	0.28	-0.27	0.30
	FP	0.21	-1.93	0.52	0.53	0.01	0.43	-0.11	0.01	0.31	0.31	0.21	0.22
YOLOX	FN	0.13	0.53	-0.07	0.05	0.56	0.00	0.17	0.20	0.21	0.41	-0.16	0.47
	FP	0.31	-1.48	0.62	0.44	0.02	0.46	0.03	0.14	0.31	0.24	0.21	0.26
CenterNet	FN	0.30	0.44	-0.05	0.11	0.33	0.06	0.14	0.14	0.15	0.35	-0.02	0.30
	FP	-0.16	-1.94	0.75	0.42	-0.04	0.04	0.04	-0.17	-0.07	0.17	0.09	0.36
GT	Train	-0.05	-0.24	0.57	0.32	0.09	-0.21	0.04	0.03	0.13	0.31	0.44	0.11
CR-	FN	0.38	0.74	-0.01	-0.09	-0.14	-0.09	0.23		0.31	0.46	0.04	0.36
CNN	FP	0.52	-2.76	0.65	0.67	0.50	0.18	-0.07	-0.04	0.13	0.16	0.69	0.22
FSAF	FN	0.33	0.71	-0.00	-0.09	-0.06	-0.17	0.23	-0.07	0.27	0.34	0.08	0.38
	FP	0.63	-2.59	0.66	0.26	0.40	0.03	-0.09	0.38	0.02	0.10	0.72	0.48
DDETR	FN	0.33	0.77	-0.00	-0.18	0.03	-0.26	0.25	-0.51	0.21	0.45	0.08	0.14
	FP	0.18	-2.85	0.67	0.61	0.03	0.45	0.13	0.39	0.25	0.32	0.69	0.33
FRCNNw/ST	FN	-0.99	0.79	-0.02	0.05	-0.97	0.10	0.24	0.67	0.39	0.63	0.13	-0.38
	FP	0.39	-2.57	0.66	0.58	0.28	0.44	-0.67	0.05	0.28	0.25	0.69	0.34
YOLOX	FN	0.07	0.77	-0.05	-0.07	-0.97	0.21	0.26	-0.21	0.33	0.37	0.22	0.00
	FP	0.32	-2.08	0.62	0.76	-0.06	0.39	-0.53	0.49	0.27	0.36	0.71	0.23
CenterNet	FN	0.43	0.66	-0.01	-0.09	0.03	-0.04	0.18	-0.34	0.21	0.35	0.05	0.53
	FP	0.48	-2.57	0.66	-0.19	0.24	0.31	-0.46	0.37	0.30	0.20	0.67	0.43

4. Conclusion

In this work, we statistically evaluate three X-ray security imagery datasets, namely OPIXray [40], SIXray [26] and PIDray [41]. The performance of six contemporary detectors operating with different deep learning paradigms is also evaluated, finding that Vision-Transformers-based detectors are the most reliable detectors and, conversely, one-stage anchor-free detectors have the worst performance, especially for heavily occluded objects. In addition, an analysis of the distribution of the properties of false positives and false negatives shows a bias towards smaller mismatches and undetected instances. It is also found that small categories, such as bullets, may be predicted with unrealistic sizes leading to lower overall detection performance. These results emphasize the importance of X-ray security image benchmark dataset analysis as a factor in the improvement of current and future object detectors in this context.

References

[1] Samet Akcay and Toby P. Breckon. Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. *Pattern Recognition*, 122, February 2022. 1, 2

[2] 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 *Proc. Int. Conf. on Image Processing*, pages 1057–1061, 2016. 1, 2

[3] Samet Akcay, Mikolaj E. Kundegorski, Chris G. Willcocks, and T. Breckon. On using deep convolutional neural network architectures for automated object detection and classification within x-ray baggage security imagery. *IEEE Transactions on Information Forensics & Security*, 13(9):2203–2215, September 2018. 1, 2

[4] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019. 2

[5] 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 *Proc. Int. Conf. on Image Processing*, pages 1224–1228. IEEE, September 2021. 2

[6] William Cai, Ro Encarnacion, Bobbie Chern, Sam Corbett-Davies, Miranda Bogen, Stevie Bergman, and Sharad Goel. Adaptive sampling strategies to construct equitable training datasets. In *Proc. of the Conf. on Fairness, Accountability, and Transparency*, pages 1467–1478, 2022. 2

[7] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 3, 4

[8] 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. 1

[9] 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. 4

[10] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *Int. Journal of Computer Vision*, 111(1):98–136, Jan. 2015. 2

[11] Yona Falinie A Gaus, Neelanjan Bhowmik, Samet Akcay, and Toby Breckon. Evaluating the transferability and adversarial discrimination of convolutional neural networks for threat object detection and classification within x-ray security imagery. In *Proc. Int. Conf. On Machine Learning And Applications*, pages 420–425. IEEE, 2019. 1

[12] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1, 3, 4

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1, 2

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*, page 580–587. IEEE Computer Society, 2014. 3

[15] 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. *IEEE Transactions on Industrial Electronics*, 68(10):10248–10257, 2021. 1

[16] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *Proc. European Conf. on Computer Vision*, pages 158–171. Springer, 2012. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *Proc. Int. Conf. on Learning Representations*, 2015. 4

[18] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *Proc. Int. Conf. on Machine Learning*, pages 1078–1088. Pmlr, 2020. 2

[19] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*, pages 9572–9581, 2019. 2

[20] Yi Li and Nuno Vasconcelos. Background data resampling for outlier-aware classification. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 13218–13227, 2020. 2

[21] 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 *Computer Vision – ECCV 2014*, pages 740–755, 2014. 1, 2, 4, 5

[22] Zhongqiu Liu, Jianchao Li, Yuan Shu, and Dongping Zhang. Detection and recognition of security detection object based on yolo9000. In *Proc. Int. Conf. on Systems and Informatics*, pages 278–282. IEEE, 2018. 1

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. Int. Conf. on Computer Vision*, pages 10012–10022, 2021. 1, 3, 4

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[25] Domingo Mery, Daniel Saavedra, and Mukesh Prasad. X-ray baggage inspection with computer vision: A survey. *IEEE Access*, 8:145620–145633, 2020. 1, 2

[26] 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 *Proc. on Computer Vision and Pattern Recognition*, pages 2114–2123, 2019. 1, 2, 3, 5, 8

[27] Jean Ponce, Tamara L Berg, Mark Everingham, David A Forsyth, Martial Hebert, Svetlana Lazebnik, Marcin Marszalek, Cordelia Schmid, Bryan C Russell, Antonio Torralba, et al. Dataset issues in object recognition. *Toward category-level object recognition*, pages 29–48, 2006. 2

[28] Mehdi Rafiei, Jenni Raitoharju, and Alexandros Iosifidis. Computer vision on x-ray data in industrial production and security applications: A comprehensive survey. *IEEE Access*, 11:2445–2477, 2023. 1, 2

[29] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3

[31] Vladimir Riffo, Hans Lobel, and Domingo Mery. Gdxray: The database of x-ray images for nondestructive testing. *Journal of Nondestructive Evaluation*, 05 2015. 2

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. Journal of Computer Vision*, 115(3):211–252, 2015. 1, 2

[33] ECAC Secretariat. Cep implements testing of apids and evd, extending the programme to nine categories of security equipment. Technical report, European Civil Aviation Conference, February 2023. 1

[34] Bo Song, Runqing Li, Xuguang Pan, Xianglan Liu, and Yan Xu. Improved yolov5 detection algorithm of contraband in x-ray security inspection image. In *Proc. Int. Conf. on Pattern Recognition and Artificial Intelligence (PRAI)*, pages 169–174, 2022. 5

[35] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 498–512, 2018. 2

[36] 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. 1

[37] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. *Domain Adaptation in Computer Vision Applications*, pages 37–55, 2017. 2

[38] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE, 2011. 2

[39] Thomas W. Webb, Neelanjan Bhowmik, Yona Falinie Abdul Gaus, and T. Breckon. Operationalizing convolutional neural network architectures for prohibited object detection in x-ray imagery. *2021 20th IEEE International Conf. on Machine Learning and Applications (ICMLA)*, pages 610–615, 2021. 1, 2

[40] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and deocclusion attention module. In *Proc. Int. Conf. on Multimedia*, MM '20, page 138–146, 2020. 1, 2, 3, 5, 8

[41] Libo Zhang, Lutao Jiang, Ruyi Ji, and Heng Fan. Pidray: A large-scale x-ray benchmark for real-world prohibited item detection. *arXiv preprint arXiv:2211.10763*, 2022. 1, 2, 3, 5, 8

[42] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. 2019. 1, 3, 4

[43] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*, pages 840–849, 2019. 1

[44] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*, pages 840–849, 2019. 3, 4

[45] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *Proc. Int. Conf. on Learning Representations*, 2021. 3, 4

[46] Zhuotun Zhu, Lingxi Xie, and Alan L Yuille. Object recognition with and without objects. *arXiv preprint arXiv:1611.06596*, 2016. 2