Evaluating the Transferability and Adversarial Discrimination of Convolutional Neural Networks for Threat Object Detection and Classification within X-Ray Security Imagery

Yona Falinie A. Gaus[1], Neelanjan Bhowmik[1], Samet Akçay[1], Toby P. Breckon[1,2]
Department of {Computer Science[1] | Engineering[2]}, Durham University, UK

***Abstract*—X-ray imagery security screening is essential to maintaining transport security against a varying profile of threat or prohibited items. Particular interest lies in the automatic detection and classification of weapons such as firearms and knives within complex and cluttered X-ray security imagery. Here, we address this problem by exploring various end-to-end object detection Convolutional Neural Network (CNN) architectures. We evaluate several leading variants spanning the Faster R-CNN, Mask R-CNN, and RetinaNet architectures to explore the transferability of such models between varying X-ray scanners with differing imaging geometries, image resolutions and material colour profiles. Whilst the limited availability of X-ray threat imagery can pose a challenge, we employ a transfer learning approach to evaluate whether such inter-scanner generalisation may exist over a multiple class detection problem. Overall, we achieve maximal detection performance using a Faster R-CNN architecture with a ResNet_{101} classification network, obtaining 0.88 and 0.86 of mean Average Precision (mAP) for a three-class and two class item from varying X-ray imaging sources. Our results exhibit a remarkable degree of generalisability in terms of cross-scanner performance (mAP: 0.87, firearm detection: 0.94 AP). In addition, we examine the inherent adversarial discriminative capability of such networks using a specifically generated adversarial dataset for firearms detection - with a variable low false positive, as low as 5%, this shows both the challenge and promise of such threat detection within X-ray security imagery.**

***Index Terms*—X-ray imagery, deep convolutional neural networks, object detection, classification, transferability.**

I. INTRODUCTION

X-ray security screening is widely used to maintain aviation, border, and transport security. To facilitate effective screening, threat detection via scanned X-ray imagery is increasingly employed to provide a non-intrusive, internal view of scanned baggage, freight, and postal items, as illustrated in Fig. 1. This produces colour-mapped X-ray images which correspond to the material properties detected via the dual-energy X-ray scanning process. Within this context, the term *threat* refers to a prohibited item such as firearms, bladed weapons, or concealed explosives, etc. In recent years, the rapid development of deep learning has brought new insight to the automation of this X-ray imagery screening task [1], [2], where the primary task is both to localise and classify the prohibited item as it appears in the image. Therefore, in this paper, we extend the current trend of using end-to-end deep learning architectures

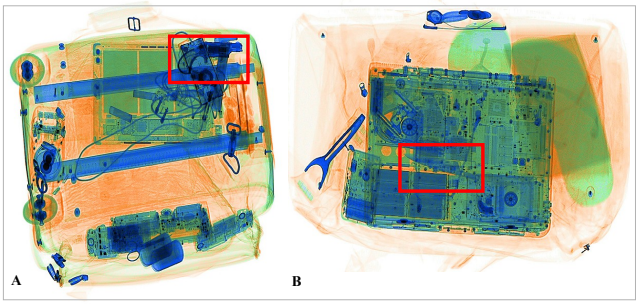

Fig. 1. Exemplar X-ray baggage imagery with prohibited items inside (red box): (A) *Firearm* and (B) *Knife*.

used for this task by performing an extended evaluation of such frameworks on large-scale X-ray security imagery. Denoted as *Dbf3* [2] and *SIXray* [3] datasets from varying X-ray scanners, we aim to provide an insight into baseline performance for several CNN architectural variants following the work of [2]. For this study, we limit our discussion to the detection of firearms (i.e. firearms and firearms with additional parts) and sharp objects (i.e. knives) as the prohibited objects. In addition, we consider a third dataset, denoted *DAD*, to specifically investigate the discriminative capability of such networks against generic *adversarial* objects, manufactured to have global shape properties similar to that of a prohibited item whilst remaining benign in nature. Subsequently, the main contributions of this paper are:

- an exploration of three end-to-end CNN-based object detection architectures with varying network configuration for addressing prohibited item detection in X-ray imagery security, expanding the work of [2]–[4].
- an evaluation of the inter-scanner transferability of such trained CNN models in terms of their generalizaton across varying X-ray scanner characteristics.
- an appraisal the trained CNN models for prohibited item discrimination against a dataset of specific adversarial objects, whose global shape characteristics closely resemble those of a firearm within X-ray imagery.

II. RELATED WORKS

In this section, scope of the literature review here is limited to prohibited item classification and detection, presented in the

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00079

following subsections:

Object Classification in X-ray Security Imagery: Early work within X-ray security imagery primarily utilises handcrafted features, where a bag of visual words (BoVW) and Support Vector Machine (SVM) are applied for feature extraction and classification, respectively [5] [6]. Mery *et al.* [7] propose a method to recognise prohibited items in multiple view X-ray imagery by filtering out false positive from monocular detection performed on single views, then match it with multiple views. A BoVW approach is further employed in [6] by exploring various feature point descriptors as visual word variants within a BoVW model achieving 94.0% accuracy for two-class firearm detection with SVM classification. The work of [1] first introduce the use of CNN to address object classification task by comparing varying CNN architectures to the earlier work extensive BoVW of [6]. Leveraging the use of transfer learning, [1] shows that CNN architectures outperform BoVW features, by achieving 98.92% detection accuracy in firearm classification. Following [1], Mery *et al.* [4] compares handcrafted features BoVW, sparse representation, codebooks with deep learning features. Consistent with the results in [1], deep features achieve higher results with more than 95% accuracy in the detection of a threat. More recently, the work on [2] exhaustively compares various CNN architectures to evaluate the impact of network complexity on overall performance. Fine tuning the entire network architecture for this problem domain yields 0.996% true positive, 0.011% false positive and 0.994% accuracy for prohibited item detection.

Object Detection in X-ray Security Imagery: Extensive experiments on object detection is conducted by Franzel *et al.* [8], where they adapt appearance-based object class detection in multiple view X-ray imagery. Multi-view detection is shown to provide superior detection performance compared with single-view detection for handguns, with mAP of 0.645. With the recent development of object detection approach, [2] examines the relative performance of traditional sliding window [8], [9] against contemporary region-based CNN variants in X-ray security imagery [10]–[13]. The work of [2] reports the performance of a traditional sliding window driven CNN detection model based on [1] against contemporary region-based and single forward-pass based CNN variants such as Faster R-CNN [10], R-FCN [11], and YOLOv2 [12] achieving a maximal 0.885 and 0.974 mAP over 6-class object detection and 2-class firearm detection problems respectively. Overall, [2] illustrates the real-time applicability and superiority of such integrated region based detection models within an X-ray security imagery context. Here we follow up on this theme, with our evaluation of the generalisation of such models by evaluating their inter-scanner transferability and discriminative capability against specific physical adversarial objects.

III. Proposed Approach

We extend the capability of contemporary region based CNN variants by incorporating Faster R-CNN [10], Mask R-CNN [14] and RetinaNet [15] as our prohibited item detection approach.

Faster R-CNN: Prohibited item detection within X-ray security imagery using Faster R-CNN were first introduced in [2], where it is trained with various network architectures such as AlexNet [16], VGG [17] and ResNet [18]. By adding a unique region proposal network (RPN) on top of the original Fast R-CNN architecture [13], it manages efficient prediction of object bounding box localization.

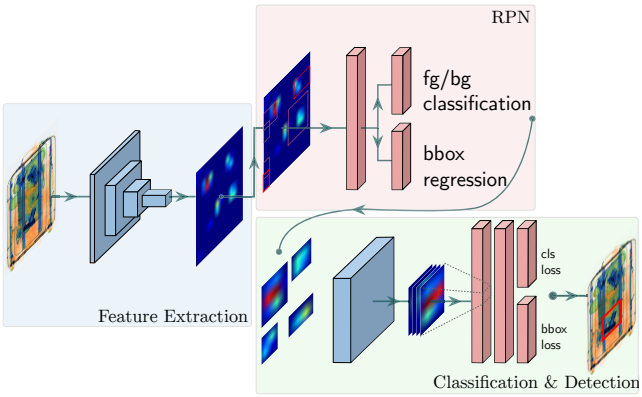

Fig. 2. Common End-to-End CNN based prohibited item detection architecture with RPN backbone.

Mask R-CNN: Following the work in [2], which builds upon Faster R-CNN in X-ray security imagery, we augment this model by adding convolutional layers to construct an object boundary segmentation mask, following the Mask R-CNN concept of [14]. It is performed by adding an additional branch to Faster R-CNN that outputs an additional image mask indicating pixel membership of a given detected object. Mask R-CNN also addresses feature map misalignment, found in Faster R-CNN [10] for higher resolution feature map boundaries, via bi-linear boundary interpolation.

RetinaNet: RetinaNet is based on a single stage detector, which involves Focal-Loss to address class imbalance issue caused by extreme foreground-background ratio [15]. In terms of X-ray security imagery, the task of identifying small metal prohibited items such as a knives presents a notable challenge due to both their size characteristics and shape overlap within the general clutter of X-ray security imagery itself. Therefore, RetinaNet is considered as it offers faster processing speed and higher accuracy for small object detection, when compared to YOLO [12], thanks its unique Feature Pyramid Network (FPN) and Focal Loss function characteristics.

Within the X-ray imagery security domain, data may be sourced from varying equipment, with different imaging parameters, X-ray energy spectra and spatial resolution [19], [20], [21]. Related work on the transferability of trained CNN models between varying X-ray scanner equipment is addressed by the work of [22], in which they focus on transfer learning between two extremities of the X-ray screening domain in terms of scale - cargo and parcel scanning (which use very different X-ray scanner equipment due to the differences in scale). From the work of [22], the two key issues identified in transferring knowledge across such X-ray domains are: *(a)* the limited availability of object of interest (prohibited item) examples, and *(b)* X-ray threat images appear in a different

machines with very different imaging characteristcs. In this work we address a similar transferability problem between X-ray scanner equipment but within the same domain (and scale) of baggage/parcel X-ray security screening.

Our hypothesis is that a CNN model trained on a given X-ray security image dataset, gathered solely from a discrete X-ray scanner in terms of manufacturer/model, will be capable of exhibiting a high degree of generalisation in terms of performance to other such datasets gathered from varying X-ray scanner configuration (varying manufacturers/models). Here, we focus on two datasets which come from different X-ray scanners. Denoted as *Dbf3* [2] and *SIXray* [3], the former dataset is from a Smith Detection X-ray scanner [21] whilst the later comes from a Nuctech scanner [20]. In addition, *Dbf3* is focused solely upon passenger carry-on baggage within an aviation security context whilst *SIXray* is based on security screening within a metro transit system context. In order to investigate cross-scanner generalisation and hence transferability, our set of CNN end-to-end models are trained using on each dataset in turn and then evaluated on the other.

IV. Experimental Setup

Our experimental setup comprises of three different datasets and a common CNN training environment. As for training details, we follow same environment as in [23].

A. X-ray Image Datasets

Our evaluation comprises three varying X-ray security imagery datasets:

Dbf3. The X-ray security imagery from Durham Dataset Full Three-class (*Dbf3*) are generated using a Smith Detection dual-energy X-ray scanner. This dataset was generated using three types metallic prohibited items, where it consists of 3,192 images of *firearms*, 1,204 images of *firearms parts*, 3,207 images of *knives*. Each object are emplaced in a representative and varied set of test bags which cover the full dimensions of aviation cabin baggage (Fig. 3A).

SIXray10. We use SIXray dataset [3] for prohibited item discovery in X-ray security images. It consists of $1,059,231$ X-ray images, in which six classes of $8,929$ prohibited items. These images are collected using a Nuctech dual-energy X-ray scanner, where the distribution of the general baggage/parcel items corresponds to stream-of-commerce occurrence. We use a subset of the SIXray dataset, *SIXray10*, which consists of five classes of prohibited items. In our experiments, we incorporate 5,083 images from two classes, 3,130 images of *firearms* and 1,953 images of *knives*, depicted in Fig. 3B.

DAD. Durham Adversarial Dataset (*DAD*) is constructed using Gilardoni dual-energy X-ray scanner (FEP ME 640 AMX), in the same manner as *Dbf3* but with artificially manufactured imitation objects, that have global shape characteristics similar to a firearm emplaced into various baggage items. These adversarial discriminative objects are L-shaped metal objects that within X-ray imagery may resemble a firearm as depicted in Figs. 3C(1) $\rightarrow$ 3C(2). This dataset consists of 200 images of imitation (adversarial) prohibited items and 200 images of real prohibited items $\{$*firearms, knifes*$\}$. This dataset is created for evaluation purposes, to test the discriminative capability of the trained CNN model as to whether it can distinguish between real and imitation of prohibited items.

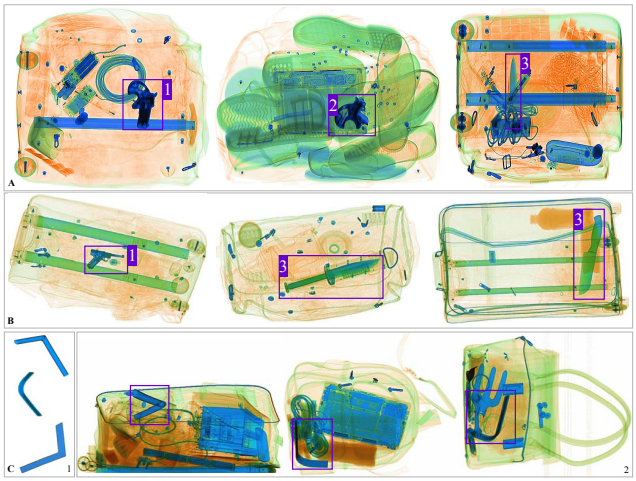

Fig. 3. Exemplar X-ray baggage images with threat objects in the purple box (1. Firearm, 2. Firearm parts, 3. Knife) from dataset (A) *Dbf3* and (B) *SIXray10*. In dataset *DAD* (C), the adversarial non-threat objects (C(1)), such as different shape metal objects (in purple box), are placed inside bags (C(2)).

V. Evaluation

In our evaluation, we consider two tasks related to prohibited item detection: *(a)* model generalisation across varying X-ray imagery datasets (Section V-B) and *(b)* model discrimination for adversarial threat objects that exhibit similar shape characteristics to the real threat objects in the same dataset.

A. Evaluation Criteria

For detection, the performance of the models is evaluated by mean average precision (mAP), as used in the seminal object detection benchmark work of [24]. We consider mean Average Precision (mAP) as our evaluation criteria following [23]. For the classification, our model performances are evaluated in terms of Accuracy (A), Precision (P), Recall (R), F-score (F1%), True Positive (TP%), and False Positive (FP%) which are calculated by thresholding ($\geqslant 0.7$) the intersection over union for detection.

B. Performance

Performance on the prohibited item detection task is carried out by comparing performance benchmark for these CNN models against prior work [2]. Here, we present a set of intra-domain results (Section V-B1, images from the same scanner dataset used for training and evaluation) against our transferablity evaluation inter-domain results (Section V-B2, images from differing scanner datasets used for training and evaluation).

1) Intra-domain Results: To provide reference performance measures, our CNN models are firstly trained and evaluated on the same dataset (i.e. *Dbf3* $\Rightarrow$ *Dbf3* and *SIXray10* $\Rightarrow$ *SIXray10*) in this set of experiments. Table I shows prohibited item detection results for Faster R-CNN [10], Mask R-CNN [14] and RetinaNet [15] with varying network configurations of ResNet. We observe that the best performance (mAP =

TABLE I

DETECTION RESULTS CNN MODELS ON TWO DATASETS. UPPER: TRAINED AND EVALUATED ON *Dbf3* WITH THREE CLASSES. LOWER: TRAINED AND EVALUATED ON *SIXray10* WITH TWO CLASSES. CLASS NAME REFLECTS CORRESPONDING AVERAGE PRECISION (AP) FOR THE INDIVIDUAL OBJECT CLASS AND MAP IS THE MEAN AVERAGE PRECISION ACROSS ALL OBJECT CLASSES.

Train ⇒ Evaluation	Model	Network configuration	Average precision			mAP
			Firearm	Firearm Parts	Knives	
Dbf3 ⇒ *Dbf3*	Faster R-CNN [10]	$ResNet_{50}$	0.87	0.84	0.76	0.82
		$ResNet_{101}$	**0.91**	**0.88**	**0.85**	**0.88**
	Mask R-CNN [14]	$ResNet_{50}$	0.86	0.83	0.75	0.81
		$ResNet_{101}$	0.89	0.86	0.80	0.85
	RetinaNet [15]	$ResNet_{50}$	0.88	0.86	0.73	0.82
		$ResNet_{101}$	0.89	0.86	0.73	0.83
SIXray10 ⇒ *SIXray10*	Faster R-CNN [10]	$ResNet_{50}$	0.87	–	0.77	0.82
		$ResNet_{101}$	0.91	–	**0.81**	**0.86**
	Mask R-CNN [14]	$ResNet_{50}$	0.87	–	0.77	0.82
		$ResNet_{101}$	0.89	–	0.79	0.84
	RetinaNet [15]	$ResNet_{50}$	0.91	–	0.79	0.85
		$ResNet_{101}$	**0.92**	–	0.79	0.86

TABLE II

DETECTION RESULTS OF CNN MODELS ON INTER-DOMAIN DATASETS. UPPER: MODELS ARE TRAINED ON *Dbf3* AND EVALUATED ON *SIXray10*. LOWER: MODELS ARE TRAINED ON *SIXray10* AND EVALUATED ON *Dbf3*.

Train ⇒ Evaluation	Model	Network configuration	Average precision		mAP
			Firearm	Knives	
Dbf3 ⇒ *SIXray10*	Faster R-CNN [10]	$ResNet_{101}$	**0.89**	**0.80**	**0.85**
	Mask R-CNN [14]	$ResNet_{101}$	0.85	0.77	0.81
	RetinaNet [15]	$ResNet_{101}$	**0.89**	0.77	0.83
SIXray10 ⇒ *Dbf3*	Faster R-CNN [10]	$ResNet_{101}$	**0.94**	**0.88**	**0.91**
	Mask R-CNN [14]	$ResNet_{101}$	0.86	0.72	0.79
	RetinaNet [15]	$ResNet_{101}$	0.87	0.66	0.76

0.88) is achieved on *Dbf3* by Faster R-CNN with $ResNet_{101}$ configuration, as presented in the upper part of Table I. Although Mask R-CNN and RetinaNet perform reasonably well for class *firearm* and *firearm parts*, these models perform less well on the *knives* class.

The *firearm parts* class is absent from the *SIXray10* dataset (as denoted in Table I). On the remaining two classes in the *SIXray10* dataset, the Faster R-CNN with $RestNet_{101}$ configuration outperformed other configuration with mAP = 0.86. As reported in the work of [3], the highest achieved AP for the class *firearm*, is 90.64% with $ResNet_{50}$. However, our model, RetinaNet with $RestNet_{101}$, produces a marginally superior AP = 0.92. The mAP results obtained for firearm detection in general are in-line with those reported in the work of [2]. Overall, the class *knives* does not perform well and this is likely to be attributable to data imbalance in the image set used for training in addition to the greater semantic difficulty in separating this item from the background clutter.

2) Inter-domain Results: This evaluation is to assess the CNN model performance across the X-ray security imagery from differing scanner sources. We use the *Dbf3* and *SIXray10* datasets from varying X-ray scanners (as described in Section IV-A). The models are trained on one dataset and evaluated on other dataset, within which the X-Ray images are generated from a different X-ray scanner (i.e. with differing energy, geometry, resolution and colour profiles). Two sets of experiments are carried out:- firstly the models are trained using *Dbf3* and evaluated on the imagery from *SIXray10* (Table II - *Dbf3* ⇒ *SIXray10*) and secondly the inverse configuration, (i.e. *SIXray10* is used for training and images from *Dbf3* for evaluation, Table II - *SIXray10* ⇒ *Dbf3*). AP/mAP is used for performance measurement for comparison as shown in Table II. We observe that for configuration *Dbf3* ⇒ *SIXray10*, a maximal mAP of 0.85 is achieved with Faster R-CNN, as presented in the upper part of Table II. Although, RetinaNet performs equally promising for the *firearm* class, for the other class (*knives*) and globally, Faster R-CNN produces superior accuracy. In general, AP of the class *knives* suffers across the models due to the variation in the shape of these objects between the datasets (Figs. 3A/B). Detection results on the *SIXray10* dataset, model training performed on the *Dbf3* dataset (i.e *Dbf3* ⇒ *SIXray10*) are depicted in Fig. 4A. Results for the reverse configuration, *SIXray10* ⇒ *Dbf3*, are presented in the lower part of Table II where we observe the maximally performing model is Faster R-CNN with 0.91 mAP and the best AP (= 0.94) for the *firearm* class. Detection results on

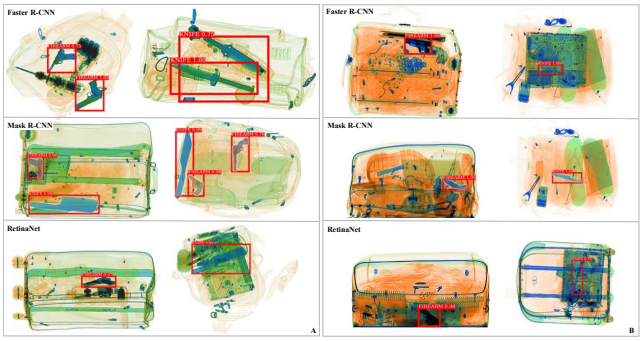

Fig. 4. Exemplar of multiple prohibited item detection for the inter-domain X-ray security training and evaluation configurations: (A) *Dbf3* ⇒ *SIXray10* and (B) *SIXray10* ⇒ *Dbf3* with varying CNN models.

TABLE III
STATISTICAL EVALUATION OF VARYING CNN ARCHITECTURES ON *Dbf3* AND *SIXray10* DATASETS (AVERAGED ACROSS ALL PROHIBITED ITEM CLASSES).

Model	Network configuration	*Dbf3* ⇒ *Dbf3*					*SIXray10* ⇒ *SIXray10*				
		A	P	F1	TP	FP	A	P	F1	TP	FP
Faster R-CNN [10]	ResNet_{50}	99.87	100.00	99.80	99.60	0.00	99.07	99.68	99.12	98.57	0.36
	ResNet_{101}	**99.96**	**100.00**	**99.93**	99.87	**0.00**	**99.83**	99.68	**99.84**	**100.00**	0.36
Mask R-CNN [14]	ResNet_{50}	99.94	99.82	99.91	**100.00**	0.09	98.65	99.68	98.72	97.78	0.36
	ResNet_{101}	99.93	99.78	99.89	100.00	0.11	99.66	99.68	99.68	99.68	0.36
RetinaNet [15]	ResNet_{50}	97.20	100.00	95.62	91.60	0.00	90.88	**100.00**	90.62	82.86	**0.00**
	ResNet_{101}	97.25	100.00	95.69	91.74	0.00	90.96	99.81	90.74	83.17	0.18

TABLE IV
STATISTICAL EVALUATION OF VARYING CNN ARCHITECTURE FOR *non-threat vs threat* CLASSIFICATION ON *DAD* DATASET.

Model	Network configuration	*Dbf3* ⇒ *DAD*					*SIXray10* ⇒ *DAD*				
		A	P	F1	TP	FP	A	P	F1	TP	FP
Faster R-CNN [10]	ResNet_{50}	82.20	79.53	82.41	**85.50**	20.95	**82.63**	87.01	**80.24**	**74.44**	10.00
	ResNet_{101}	**84.75**	**87.57**	**84.16**	81.00	**11.50**	76.75	86.39	73.20	63.50	10.00
Mask R-CNN [14]	ResNet_{50}	77.75	80.33	76.76	73.50	18.00	76.50	**89.55**	71.86	60.00	07.00
	ResNet_{101}	83.75	86.49	83.12	80.00	12.50	78.50	83.93	76.63	70.50	13.50
RetinaNet [15]	ResNet_{50}	78.22	80.11	77.20	74.50	18.14	68.75	88.66	57.91	43.00	05.50
	ResNet_{101}	79.73	82.05	81.01	80.00	20.59	67.75	89.01	55.67	40.50	**05.00**

the *Dbf3* dataset, whilst model training is performed on the *SIXray10* dataset are depicted in Fig. 4A. As anticipated, the *knives* class suffers from relatively low AP for both Mask R-CNN and RetinaNet due to the variation in visual appearance in between training and evaluation sets for this particular class (Fig. 4 A/B). In the training data, the knives are mostly placed on/under electronic items; however, the evaluation set consists of very differing shapes of knives across a diverse background. Overall, CNN models trained with the *SIXray10* dataset offers superior performance when compared to when the models are trained with *Dbf3* - even when evaluated on *Dbf3*. As a result, we can infer that although images are from differing X-ray scanners, the transferability of learnt CNN models is viable in terms of maintaining prohibited item detection performance over varying X-ray imagery sources.

C. Adversarial Discriminative Objects

Furthermore, we evaluate the discriminative capability of the CNN-based detection models we consider (Section III), trained for multiple class object detection (as per Table I, when tested against both real threat objects and imitation (adversarial) non-threat objects that have the same global shape and material characteristics as the real threat objects. Our test dataset for this task, *DAD*, is fully described in the Section IV-A. To provide an initial benchmark for performance without such adversarial examples, detection results for the three-class prohibited item problem with the *Dbf3* dataset and two-class threat problem within the *Sixray10* dataset, averaged across all object classes, are presented in Table III (calculated as per Section V-A). Here we can observe performance such that all of the models considered consistently offer very low false positive (FP) complimented by a high true positive (TP) detection across both problems (see Table III).

To establish the impact of introducing adversarial examples, we make use of the *DAD* dataset (Section IV-A) containing our imitation (adversarial) threat-like objects, constructed as a series of simple L-shaped metal brackets, mimicking the real shape of *firearms*, *firearm parts* or *knives* within X-ray security imagery depending on the angle of view (see Fig. 3C). In order to illustrate the impact of these examples, on overall detection performance we introduce a global *'threat vs. non-threat'* detection problem on the basis that the *DAD* dataset (Section IV-A) has a $50/50$ split between X-ray security images containing a genuine *threat* object belonging to the set {*firearm, firearm parts, knives*} and benign (*non-threat*) images containing our imitation (adversarial) threat-like objects. All genuine *threat* and (*non-threat*) adversarial objects are set amongst regular benign baggage clutter. This gives rise to a simple two-class meta-problem of *'threat vs. non-threat'* by combining true positive detection for any of the set {*firearm, firearm parts, knives*} in the genuine *threat* object images as the class *threat* and conversely defining false positives as detection for any of these objects within the benign (*non-threat*) images that have the imitation (adversarial) threat-like objects present.

As per the results of the performance benchmark shown in Table III, we evaluate the same CNN models trained on each of the *Dbf3* and *SIXray10* datasets and evaluate on *DAD* dataset for this two-class meta-problem, {*threat, non-threat*} (Table IV). With a *Dbf3* trained model, Faster R-CNN with ResNet_{101} achieves maximal performance with the lowest FP (11.50%) and accuracy of (84.75%) (Table IV). However, Faster R-CNN with ResNet_{50} has maximal TP (85.50%), yet significantly higher FP. Conversely, the lowest FP (5%) is achieved by RetinaNet with ResNet_{101} with a *SIXray10* trained model but this model suffers from very low TP (40.5%). The Mask R-CNN produces 7% FP with reasonable accuracy of 78.5%. By comparing the performance of these models, under both standard conditions (benchmark performance in Table III) and adverserial conditions (Table IV), we can immediately see the impact of the adversarial threat-like imitators as the models get confused by the L-shaped imitation objects and wrongly classifies them as threat objects (Fig. 5). This clearly illustrates the challenge posed by such physical adversarial object examples within achieving viable performance for automated

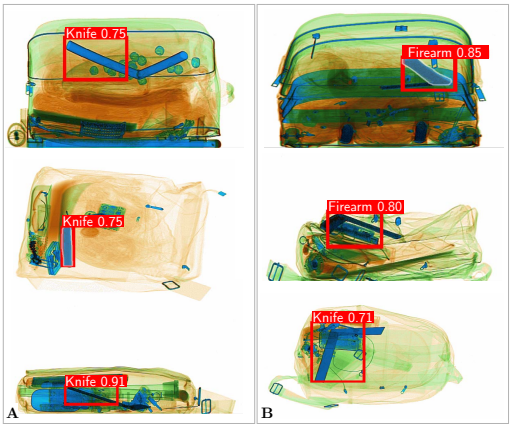

Fig. 5. Exemplar where CNN threat detection falsely detect L-shaped metal item as threat item when the models (Faster R-CNN: row.1, Mask R-CNN: row.2 and RetinaNet: row.3) are trained on (A) *Dbf3* and (B) *SIXray10* dataset.

X-ray security image classification.

VI. Conclusion

This paper explores the transferability and adversarial discrimination of various end-to-end object detection Convolutional Neural Network (CNN) architectures for prohibited item detection within X-ray security imagery. Faster R-CNN achieves superior baseline performance (0.86/0.88 mAP) over a three class prohibited item detection problem (objects: *firearms, firearm parts, knives*) evaluated on two disparate datasets capturing with varying X-ray image scanner equipment. Furthermore, we directly evaluate transferability of such CNN model performance by employing cross-scanner validation to ascertain inter-scanner generalisation performance. We show that a CNN model trained on X-ray security imagery exclusively from one X-ray scanner manufacturer's device and then performance tested exclusively on separate X-ray security imagery from another manufacturer's scanner will produce strong generalisation performance despite differences in the X-ray image characteristics (0.85/0.91 mAP, two class problem - *firearms, knives*). This provides strong insight to the generalisation capability of the proposed method across varying X-ray imagery characteristics. Finally, we appraise the performance of such trained CNN models against physically constructed adversarial examples (imitation threat items). Whilst this shows a clear impact on generalised performance from the use of such adversarial object, it additionally illustrates the possibility of a false positive rate as low as 5% remains under such conditions.

Acknowledgements: The authors would like to thank the UK Home Office for partially funding this work. Views contained within this paper are not necessarily those of the UK Home Office.

References

[1] S. Akçay, M. E. Kundegorski, M. Devereux, and T. P. Breckon, "Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery," in *International Conference on Image Processing*. IEEE, 2016, pp. 1057–1061.

[2] S. Akçay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, "Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2203–2215, 2018.

[3] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. Ye, "Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2119–2128.

[4] D. Mery, E. Svec, M. Arias, V. Riffo, J. M. Saavedra, and S. Banerjee, "Modern computer vision techniques for x-ray testing in baggage inspection," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 4, pp. 682–692, April 2017.

[5] M. Bastan, W. Byeon, and T. M. Breuel, "Object recognition in multi-view dual energy x-ray images." in *British Machine Vision Conference*, vol. 1, no. 2, 2013, p. 11.

[6] M. E. Kundegorski, S. Akcay, M. Devereux, A. Mouton, and T. P. Breckon, "On using feature descriptors as visual words for object detection within x-ray baggage security screening," in *International Conference on Imaging for Crime Detection and Prevention*, 2016, pp. 1–6.

[7] D. Mery, V. Riffo, I. Zuccar, and C. Pieringer, "Object recognition in x-ray testing using an efficient search algorithm in multiple views," *Insight - Non-Destructive Testing and Condition Monitoring*, vol. 59, no. 2, pp. 85–92, 2017.

[8] T. Franzel, U. Schmidt, and S. Roth, "Object detection in multi-view x-ray images," in *Pattern Recognition*, 2012, pp. 144–154.

[9] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *International Conference on Learning Representations, CBLS, April 2014*, 2014.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[11] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 379–387.

[12] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7263–7271.

[13] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[19] "Gilardoni- x-ray and ultrasounds," https://www.gilardoni.it/en/, accessed: 2019-10-18.

[20] "Nuctech official site," http://www.nuctech.com/en, accessed: 2019-10-18.

[21] "Smiths detection," https://www.smithsdetection.com/, accessed: 2019-10-18.

[22] M. Caldwell, M. Ransley, T. Rogers, and L. Griffin, "Transferring x-ray based automated threat detection between scanners with different energies and resolution," in *Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies*, vol. 10441. Int. Society for Optics and Photonics, 2017, p. 104410F.

[23] Y. F. A. Gaus, N. Bhowmik, S. Akçay, P. M. Guillen-Garcia, J. W. Barker, and T. P. Breckon, "Evaluation of a dual convolutional neural network architecture for object-wise anomaly detection in cluttered x-ray security imagery," in *International Joint Conference on Neural Networks*, 2019.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision*, 2014, pp. 740–755.