ON THE IMPACT OF USING X-RAY ENERGY RESPONSE IMAGERY FOR OBJECT DETECTION VIA CONVOLUTIONAL NEURAL NETWORKS

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

ABSTRACT

Automatic detection of prohibited items within complex and cluttered X-ray security imagery is essential to maintaining transport security, where prior work on automatic prohibited item detection focus primarily on pseudo-colour (*rgb*) X-ray imagery. In this work we study the impact of variant X-ray imagery, i.e., X-ray energy response (*high, low*) and effective-z compared to *rgb*, via the use of deep Convolutional Neural Networks (CNN) for the joint object detection and segmentation task posed within X-ray baggage security screening. We evaluate state-of-the-art CNN architectures (Mask R-CNN, YOLACT, CARAFE and Cascade Mask R-CNN) to explore the transferability of models trained with such 'raw' variant imagery between the varying X-ray security scanners that exhibits differing imaging geometries, image resolutions and material colour profiles. Overall, we observe maximal detection performance using CARAFE, attributable to training using combination of *rgb, high, low*, and effective-z X-ray imagery, obtaining 0.7 mean Average Precision (mAP) for a six class object detection problem. Our results also exhibit a remarkable degree of generalisation capability in terms of cross-scanner transferability (AP: $0.835/0.611$) for a one class object detection problem by combining *rgb, high, low*, and effective-z imagery.

***Index Terms*—** x-ray imagery, deep convolutional neural network, object detection, transferability

1. INTRODUCTION

X-ray security screening plays a pivotal role in aviation security. However, manual inspection of potentially prohibited items is challenging due to the clutter and occlusion present within X-ray scanned baggage. A modern X-ray security scanner makes use of multiple X-ray energy levels in order to facilitate effective materials discrimination [1]. Subsequently, a dual-energy X-ray scanner imagery consists of two intensity images acquired at two discrete energy levels (*low* and *high*), facilitating the recovery of material properties (effective atomic number, effective-z). The information is fused with the help of a colour transfer function into a single pseudo-colour X-ray image (Figure 1A) to facilitate the interpretation of the baggage contents [2].

The advancement of deep Convolutional Neural Networks (CNN) has brought new insight to the automation of this X-ray imagery screening task [3–5] where the primary task is both to localise and classify the prohibited items. Prior works [4,6] are concentrated on the shaped-based detection of prohibited items achieving both high detection performance with low false positive. The work of [3] uses a pre-trained GoogleNet model for classification task in X-ray baggage scans for detecting potentially prohibited items. Subsequently, the work of [4] compares contemporary region-based and single forward-pass based CNN architectures (Faster R-CNN [7], YoloV2 [8]) achieving 0.88 and 0.97 mean Average Precision (mAP) over six class and two class X-ray baggage security object detection problems respectively. Following these works, [5] proposes a dual-stage CNN architecture for anomaly detection in a six class problem. Semi-supervised adversarial learning is used in the works of [9, 10] for prohibited item detection. The availability of the large-scale X-ray baggage datasets (SIXray [11], and OPIXray [12]) has provided further insight into the transferability and generalisation abilities of the CNN architectures [6] across varying X-ray security scanners which all exhibit varying characteristics in terms of projection geometry common resolution and pseudo-colour mapping. While most prior work [4–6] process each view of multi-view X-ray scanners independently, [13] utilise the corresponding information between the views for a detection task achieving 0.91 mAP.

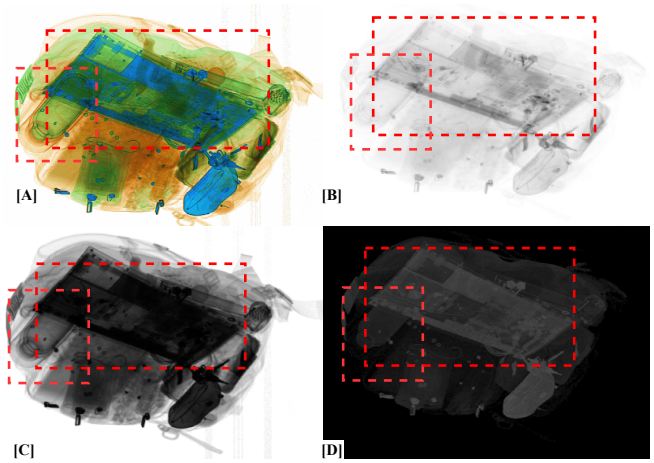

Fig. 1. Exemplar *rgb* (A), *high* (B), *low* (C), and effective-z (D) X-ray imagery from *deei6* dataset containing target classes in bounding boxes.

Almost all of the prior work discussed here only use pseudo-colour/false colour (*rgb*) X-ray imagery that is itself generated from the 'raw' *high/low*/effective-z imagery ob-

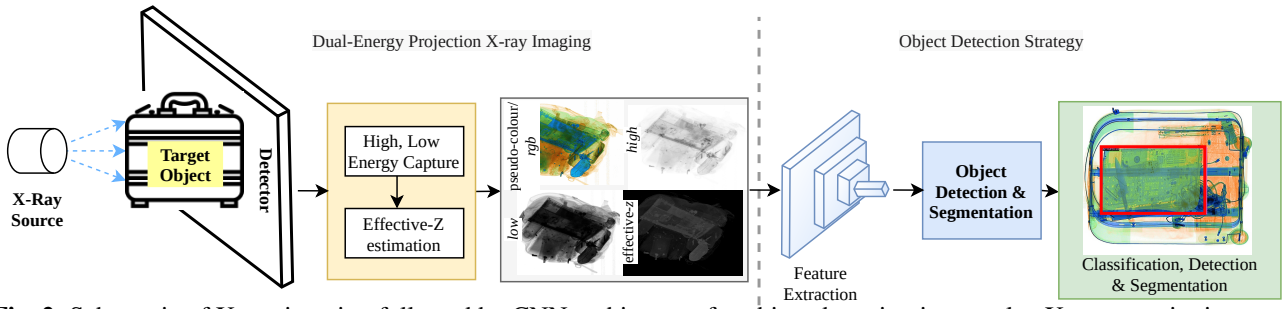

Fig. 2. Schematic of X-ray imaging followed by CNN architecture for object detection in complex X-ray security imagery.

tained from the scanner. By contrast, in this study we consider the impact of using this 'raw' imagery (Figure 1(B)→(D)) directly for the purposes of prohibited object detection. The objective of using two energy levels (*high* and *low*) for object detection task is to obtain both the density and atomic number Z (effective-z) of the scanned materials [14], as the intensity values in the energy response may encode very valuable material information, which is not as readily identifiable within the pseudo-colour X-ray imagery.

Against this background, this paper introduces the following novel contributions: (a) an experimental evaluation of dual-energy X-ray imagery for joint object detection and segmentation task, via use of characteristically diverse end-to-end CNN architectures [15–18], (b) an investigation into the inter-scanner transferability of such CNN models, trained on dual-energy X-ray imagery, in terms of their generalisation across varying X-ray scanner characteristics.

2. PROPOSED APPROACH

In this study, we present dual-energy X-ray imaging technique (Figure 2, left) in Section 2.1 and followed by object detection and segmentation strategies (Figure 2, right) in Section 2.2.

2.1. Dual-Energy Projection X-ray Imaging

The primary components of X-ray security scanner system are composed of an X-ray source emitter and detector (Figure 2, left). X-rays are emitted with photon energy ranging up to 150kV [19] from a X-ray source. Generally, the X-ray images are constructed by attenuating the signal on the material as the target object proceeds through the scanner tunnel, defined as $I(E) = I_0 e^{-\mu t}$, where $I(E)$ is the captured intensity as a function of the thickness t, the emitted intensity I_0 and the absorption coefficient μ. The absorption coefficient is defined by $\mu = \alpha(Z, E)\rho$, where Z is the atomic number, E is the energy, ρ is the density, and $\alpha(Z, E)$ corresponds to the mass attenuation coefficient in terms of Z and E [20].

In the dual-energy source X-ray imaging, two intensity responses captured at two different energy levels, *low* and *high* ($E = \{l, h\}$) and are subsequently combined to construct *low* and *high* energy response images (Figure 2, left). Given the Compton scatter coefficient (μ_c) and the photoelectric absorption coefficient (μ_p) [21], material identification (approximate atomic number, effective-z; Z_eff) can be calculated as:

$$Z\_eff = K^{'}(\frac{\mu_p}{\mu_c})^{\frac{1}{n}} \quad (1)$$

where $K^{'}$ and n are constant [21]. In this work, we evaluate the use of the pseudo-colour (*rgb*), dual-energy response (*h, l*) and effective-z (Figure 1) as alternative inputs imagery for CNN-based object detection.

2.2. Object Detection and Segmentation Strategy

We consider four contemporary CNN architectures of differing characteristics, spanning both single stage and multi stage detection approaches, and explore their applicability for prohibited item detection within varying configurations of dual-energy X-ray imagery inputs.

Mask R-CNN [15] is a two-stage detector for object instance segmentation, developed on top of Faster R-CNN [7]. Mask R-CNN [15] uses the Faster R-CNN [7] architecture for feature extraction, Region Proposal Network (RPN), and followed by region of interest alignment (RoIAlign) via bilinear boundary interpolation to produce higher resolution feature map boundaries suitable for input into a secondary classifier. The output from the RoIAlign layer is subsequently fed into a series of segmentation processing layers (mask head), that generate an additional image mask indicating pixel membership of a given detected object.

YOLACT [16] is an one-stage detector, based on RetinaNet [22], that directly predicts boxes without a separate region proposal step. YOLACT [16] generates a set of prototype masks, linear combination coefficients for each predicted instance, and associated bounding boxes. It combines the prototype masks using the corresponding predicted mask coefficients followed by cropping with a predicted bounding box to generate the final output.

CARAFE [17] is a two-stage architecture, which proposes effective feature up-sampling operators and integrates it into Feature Pyramid Network to boost the performance. For instance segmentation, a feature map, which represents the object shape accurately, is used to predict the final instance segmentation result.

Cascade Mask R-CNN [18], a multi-stage detector, is a hybrid of Cascade R-CNN and Mask R-CNN [15]. Similar to Mask R-CNN [15], each stage has a segmentation mask

branch, a label prediction branch, and a bounding box detector branch. The current stage will accept RPN or the bounding box returned by the previous stage as an input. The second stage increases localisation performance accuracy, and subsequently, it further refines the output. This is repeated over multiple stages with increasingly refined criteria for discarding low-quality proposals from the previous stage such that it predicts precise bounding boxes and masks at the final stage.

In this study, we compare these four CNN architectures for object detection (Figure 2, right) using combination of different variants dual-energy X-ray imagery (Section 2.1). To assess the impact of dual-energy X-ray imagery variants on object detection we first use *rgb*, *high* (h), *low* (l), and effective-z (z) imagery individually. Secondly, h, l and z are combined as three channels (*hlz*) images. Thirdly, we combine *rgb, high, low*, and effective-z imagery for joint object detection and segmentation task.

Within the X-ray imagery security domain, imagery may be sourced using varying scanners [19, 23, 24], which have different X-ray energy spectra, spatial resolution and material colour profiles. In prior work [6, 25] on transferability and generalisation ability, [25] focuses on transfer learning between cargo parcel scanning (different scanner equipment due to the differences in scale). The work of [6] shows cross-scanner transferability of CNN architectures (using *rgb* X-ray imagery) in terms of their generalisation across varying X-ray scanner characteristics. In this study, we further evaluate the effectiveness of using variants of dual-energy X-ray imagery (Section 2.1) on generalisation capabilities of the CNN architectures.

3. EVALUATION

We focus on three datasets that are sourced from different X-ray scanners [19, 23, 24]. The *deei6* is created from a Gilardoni X-ray scanner [19], and consists of *rgb*, *high, low*, and effective-z imagery. The other two datasets, *dbs_laptop* and *dbr_laptop*, are generated by a Smith Detection [23] and Rapiscan X-ray scanner [24] respectively and consist of *rgb* X-ray imagery. The four CNN architectures (Section 2.2) are trained using *rgb* and combinations of *rgb, high, low*, and effective-z X-ray imagery from *deei6* dataset. Subsequently, we evaluate the model performance on *rgb* X-ray imagery of *dbs_laptop* and *dbr_laptop* datasets.

deei6: Our dataset (Durham Electrical and Electronics Items) is constructed using a dual-energy Gilardoni FEP ME 640 AMX scanner [19] with associated pseudo-colour materials mapping. This dataset is composed of six-classes of consumer electronics, electrical and other items: {*bottle, hairdryer, iron, toaster, phone-tablet, laptop*}, totalling $7,022$ images (70:30 data split for experiments). We also access the *high, low*, and effective-z imagery to construct *deei6*$_{rgb}$, *deei6*$_{h}$, *deei6*$_{l}$ and *deei6*$_{z}$ imagery as depicted in Figure 1.

To investigate the generalisation capabilities of the CNN architectures, we also use the following two datasets:

dbs_laptop: comprises 488 *laptop* class *rgb* X-ray image examples (with associated pseudo-colour materials mapping), which is sourced from a Smith Detection X-ray scanner [23].

dbr_laptop: comprises 107 *laptop* class X-ray *rgb* image examples (with associated pseudo-colour materials mapping). This dataset is sourced from Rapiscan 620DV X-ray scanner [24].

The CNN architectures (Section 2.2) are implemented using MMDetection framework [26]. Through the transfer learning paradigm, training (using X-ray imagery variants) of all CNN architectures (Section 2.2) are initialised with ImageNet [27] pretrained weights (which originate from training on colour RGB imagery). Our CNN architectures are trained using ResNet$_{50}$ [28] backbone with following training configuration: backpropagation optimisation performed via Stochastic Gradient Descent, initial learning rate of 2.5×10^{-4} with decay by a factor of 10 at 7^{th} epoch, and a batch size of 4. The model performance is evaluated by MS-COCO metrics [29] (IoU of $0.50 : .05 : 0.95$), using Average Precision (AP) for class-wise and mAP for overall performance.

3.1. Impact of Dual-energy X-ray Imagery

In the first set of experiments (Table 1), exemplar items in X-ray security imagery are detected using the CNN architectures set out in Section 2.2. We use variants of dual-energy X-ray imagery of the *deei6* dataset for training and evaluation denoted as *deei6*$_x$ for $x = \{rgb, h, l, z, hlz\}$. The highlighted mAP signifies the maximal results obtained for overall performance. At first, the CNN architectures are trained and evaluated on *rgb* X-ray imagery (Table 1, *rgb*), in line with [4,5,11]. The best performance is achieved by Cascade Mask R-CNN (CM RCNN) [18] producing maximal mAP (0.693) and outperforming other three CNN architectures. When we train CNN architectures using *high*, *low*, and effective-z imagery individually and together as three channels (*hlz*), the overall performance (Table 1) does not improve compared to *rgb* imagery. The lowest performing training set is *deei6*$_z$ imagery achieving only 0.627 of mAP (with Cascade Mask R-CNN [18]). It is possibly due to the lack of contrast in the pixel intensity in effective-z imagery where the target objects appear similar to the background, leading to inferior detection performance. The impact of dual-energy X-ray imagery can be observed while combining *rgb, high, low* and effective-z (*deei6*$_{rgb,hlz}$) together. The maximal mAP of 0.7 (Table 1, *rgb,hlz*) is achieved by CARAFE [17] marginally outperforming *rgb* imagery (mAP: 0.693). Although YOLACT [16] is the simplest architecture (34.76 million parameters), it outperforms Mask R-CNN [15] while training using *rgb,hlz* X-ray imagery (mAP: 0.686 vs 0.680).

In the confusion matrices (Figure 3) of CARAFE [17], we observe strong true positive (diagonal) and low false positive (off-diagonal) occurrence. The advantage of combining *rgb, high, low*) and effective-z can be seen in the class *phone-tablet* (0.698 to 0.729, Figure 3(A)$\rightarrow$(B)), with improvement of confidence in localising small objects within cluttered X-

ray security imagery.

	Model	Bottle	Hairdryer	Iron	Toaster	P-tablet	Laptop	mAP
$deei6_{rgb}$	M RCNN	0.633	0.651	0.688	0.793	0.550	0.747	0.677
	YOLACT	0.646	0.596	0.672	0.784	0.540	0.770	0.668
	CARAFE	0.637	0.638	0.692	0.788	0.543	0.770	0.678
	CM RCNN	0.650	0.659	0.708	0.801	0.560	0.781	**0.693**
$deei6_{h}$	M RCNN	0.607	0.615	0.665	0.761	0.521	0.745	0.652
	YOLACT	0.641	0.597	0.649	0.756	0.533	0.765	0.657
	CARAFE	0.631	0.624	0.676	0.754	0.522	0.757	0.661
	CM RCNN	0.632	0.638	0.687	0.782	0.539	0.783	**0.677**
$deei6_{l}$	M RCNN	0.597	0.605	0.670	0.779	0.520	0.749	0.653
	YOLACT	0.619	0.576	0.659	0.771	0.520	0.760	0.651
	CARAFE	0.632	0.606	0.662	0.777	0.530	0.768	0.662
	CM RCNN	0.641	0.627	0.677	0.784	0.541	0.778	**0.674**
$deei6_{z}$	M RCNN	0.543	0.521	0.629	0.798	0.489	0.716	0.616
	YOLACT	0.548	0.395	0.597	0.783	0.477	0.737	0.589
	CARAFE	0.550	0.492	0.629	0.786	0.522	0.718	0.616
	CM RCNN	0.560	0.516	0.634	0.796	0.507	0.749	**0.627**
$deei6_{hlz}$	M RCNN	0.613	0.617	0.667	0.789	0.535	0.742	0.660
	YOLACT	0.615	0.575	0.644	0.757	0.525	0.756	0.645
	CARAFE	0.639	0.611	0.673	0.791	0.557	0.765	0.673
	CM RCNN	0.632	0.630	0.689	0.802	0.541	0.775	**0.678**
$deei6_{rgb,hlz}$	M RCNN	0.644	0.633	0.682	0.799	0.543	0.779	0.680
	YOLACT	0.670	0.625	0.676	0.796	0.560	0.791	0.686
	CARAFE	0.676	0.653	0.690	0.808	0.580	0.792	**0.700**
	CM RCNN	0.667	0.663	0.696	0.806	0.552	0.798	0.697

Table 1. Object detection results of CNN architectures using different X-ray imagery from the *deei6* dataset.

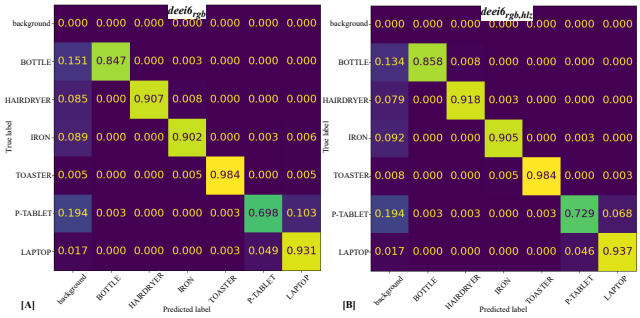

Fig. 3. Confusion Matrix of the CARAFE [17] trained on *rgb* (A) and combination of {*rgb,hlz*} (B) X-ray imagery.

3.2. Cross-scanner Transferability

In this set of experiments (Table 2), we assess the CNN architecture performance across the X-ray imagery (*dbs_laptop* and *dbr_laptop*) from different scanner sources [23, 24]. The CNN architectures are trained using variants of dual-energy X-ray imagery of *deei6* dataset but evaluated on a test set of only rgb pseudo-colour imagery (*dbs_laptop* and *dbr_laptop*). The positive impact of combining {*rgb,hlz*} X-ray imagery is evident with all four CNN architectures (Table 2). For *dbs_laptop*, CARAFE [17] produces the best performance (AP: 0.835, Table 2, lower) when trained using combination of *rgb, high, low* and effective-z X-ray imagery, significantly outperforming *rgb* X-ray imagery (AP: 0.763, Table 2, upper). Similar significant performance improvement is noticeable on *dbr_laptop* dataset with CARAFE [17] achieving the highest AP of 0.611 (Table 2, lower). A plausible explanation for the performance improvement is that the variation in X-ray imagery by combining *rgb, high, low* and effective-z imagery during training, leads the CNN architectures to learn meaningful image features, which alleviates to achieve a higher degree of model generalisation in object detection within X-ray imagery. Although CARAFE [17] is a simpler architecture (49.41 million parameters) compared to the Cascade Mask R-CNN [18] (77.04 million parameters), it offers a better generalisation ability by training on a more varied set of multiple X-ray imagery variants. In Figure 4A the target *laptop* is missed in both test images when trained solely on *rgb* imagery, but successfully detected when trained with combined {*rgb,hlz*} X-ray imagery (Figure 4B). Hence, we can deduce that although X-ray images are from differing scanners, the transferability of the trained CNN models is significantly improved by training over a more varied training set that includes both pseudo-colour *rgb* and variant dual-energy X-ray imagery.

Training Dataset	Model	Test-set	
		dbs_laptop	*dbr_laptop*
$deei6_{rgb}$	M RCNN	0.749	0.530
	YOLACT	0.633	0.344
	CARAFE	0.775	0.476
	CM RCNN	0.763	0.518
$deei6_{rgb,hlz}$	M RCNN	0.807	0.593
	YOLACT	0.782	0.521
	CARAFE	**0.835**	**0.611**
	CM RCNN	0.803	0.587

Table 2. Object detection results (AP) on *dbs_laptop* and *dbr_laptop* datasets, where CNN architectures are trained using variant X-ray imagery from the *deei6* dataset.

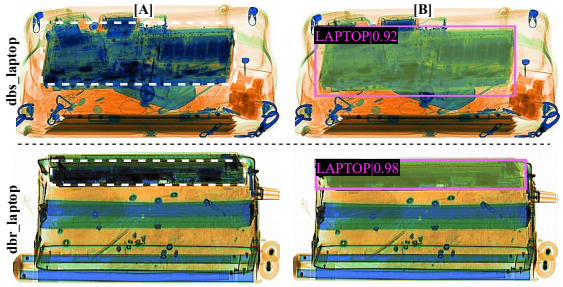

Fig. 4. Detection examples from *dbs_laptop* and *dbr_laptop* using CARAFE [17] trained on *rgb* (A) and {*rgb,hlz*} (B) X-ray imagery from *deei6* dataset. White dashed box in (A) fails to detect the target.

4. CONCLUSION

This work examines the impact of X-ray imagery variants, i.e., dual-energy X-ray responses (*high, low*), effective-z and pseudo-colour (*rgb*), via the use of CNN architectures for the object detection task posed within X-ray baggage security screening. We illustrate that the combination of *rgb, high, low* and effective-z X-ray imagery produces maximal performance across all four CNN architectures for a six classes object detection problem, with CARAFE [17] achieving the highest mAP of 0.7. Furthermore, our results also demonstrate a remarkable degree of generalisation capability in terms of cross-scanner transferability (AP: 0.835/0.611 with CARAFE [17]) for a one class object detection problem by combining {*rgb, hlz*} X-ray imagery. This clearly illustrates a strong insight into the benefits of using a combination of dual-energy X-ray imagery for object detection and segmentation tasks, which could additionally useful for component-wise anomaly detection analysis. Future work will consider the use of dual-energy variant imagery for combined material discrimination and anomaly detection within cluttered X-ray security imagery.

5. REFERENCES

[1] S. Singh and M. Singh, "Explosives detection systems (eds) for aviation security," *Signal processing*, vol. 83, no. 1, pp. 31–55, 2003.

[2] D. Turcsany, A. Mouton, and T.P. Breckon, "Improving feature-based object recognition for x-ray baggage security screening using primed visual words," in *Proc. Int. Conf. on Industrial Technology*. February 2013, pp. 1140–1145, IEEE.

[3] 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 *Proc. Int. Conf. on Image Processing*. IEEE, 2016, pp. 1057–1061.

[4] 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.

[5] 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 *Proc. Int. Joint Conf. on Neural Networks*, 2019.

[6] Y.F. A Gaus, N. Bhowmik, S. Akcay, and T.P. 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*, 2019, pp. 420–425.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[8] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 6517–6525.

[9] S. Akcay, A. Atapour-Abarghouei, and T.P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. on Computer Vision*. Springer, 2018, pp. 622–637.

[10] S. Akçay, A. Atapour-Abarghouei, and T.P. Breckon, "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection," in *Proc. Int. Joint Conf. on Neural Networks*. IEEE, 2019, pp. 1–8.

[11] 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 *Proc. Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 2119–2128.

[12] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module," in *Proc. Int. Conf. on Multimedia*, 2020, p. 138–146.

[13] B.K.S. Isaac-Medina, C.G. Willcocks, and T.P. Breckon, "Multi-view object detection using epipolar constraints within cluttered x-ray security imagery," in *Proc. Int. Conf. Pattern Recognition*. 2020, IEEE.

[14] V. Rebuffel and J.M. Dinten, "Dual-energy x-ray imaging: benefits and limits," *Insight-non-destructive testing and condition monitoring*, vol. 49, no. 10, pp. 589–594, 2007.

[15] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. Int. Conf. on Computer Vision*, 2017.

[16] D. Bolya, C. Zhou, F. Xiao, and Y.J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. of Int. Conf. on Computer Vision*, 2019.

[17] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "CARAFE: Content-aware reassembly of features," in *Proc. of Int. Conf. on Computer Vision*, 2019, pp. 3007–3016.

[18] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[19] "Gilardoni," www.gilardoni.it/en/security/x-ray-solutions/automatic-detection-of-explosives/fep-me-640-amx, Accessed: 2021-01-10.

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

[21] André Mouton and Toby P. Breckon, "A review of automated image understanding within 3d baggage computed tomography security screening.," *Journal of X-ray science and technology*, vol. 23 5, pp. 531–55, 2015.

[22] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. Int. Conf. on Computer Vision*, 2017, pp. 2999–3007.

[23] "Smiths detection," www.smithsdetection.com, Accessed: 2021-01-10.

[24] "Rapiscan," www.rapiscansystems.com/en/products/rapiscan-620dv, Accessed: 2021-01-10.

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

[26] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C.C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[27] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Prof. Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

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

[29] 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 *Proc. European Conf. on Computer Vision*, 2014, pp. 740–755.