The Good, the Bad and the Ugly: Evaluating Convolutional Neural Networks for Prohibited Item Detection Using Real and Synthetically Composited X-ray Imagery

Neelanjan Bhowmik[1]

Qian Wang[1]

Yona Falinie A. Gaus[1]

Marcin Szarek[3]

Toby P. Breckon[12]

Department of
{Computer Science[1] | Engineering[2]}
Durham University
Durham, UK

[3]School of Engineering
Cranfield University
Cranfield, UK

Abstract

Detecting prohibited items in X-ray security imagery is pivotal in maintaining border and transport security against a wide range of threat profiles. Convolutional Neural Networks (CNN) with the support of a significant volume of data have brought advancement in such automated prohibited object detection and classification. However, collating such large volumes of X-ray security imagery remains a significant challenge. This work opens up the possibility of using synthetically composed imagery, avoiding the need to collate such large volumes of hand-annotated real-world imagery. Here we investigate the difference in detection performance achieved using real and synthetic X-ray training imagery for CNN architecture detecting three exemplar prohibited items, {*Firearm, Firearm Parts, Knives*}, within cluttered and complex X-ray security baggage imagery. We achieve 0.88 of mean average precision (mAP) with a Faster R-CNN and ResNet$_{101}$ CNN architecture for this 3-class object detection using real X-ray imagery. While the performance is comparable with synthetically composited X-ray imagery (0.78 mAP), our extended evaluation demonstrates both challenge and promise of using synthetically composed images to diversify the X-ray security training imagery for automated detection algorithm training.

1 Introduction

To ensure transport and border security, X-ray security screening is commonplace within public transport and border security installations such as airports, railway and metro stations. However, due to the nature of cluttered and complex X-ray imagery (Figure 1), the process of

X-ray screening is complicated by tightly packed items within baggage making it challenging and time-consuming to identify the presence of prohibited items. With the natural occurrence of such prohibited items being rare, previous studies cite time constraints as a major factor in the performance limitations of human operators for this screening task [3, 25].

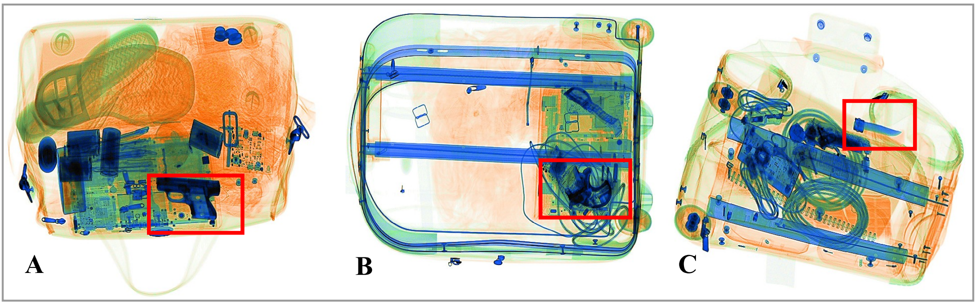

Figure 1: Exemplar X-ray security baggage images with prohibited objects - red box: (A) Firearm (B) Firearm Parts and (C) Knife.

Whilst challenging for a human, a reliable automatic prohibited item detection system may assist in improving the performance and throughput of such screening processes [27]. To date, contemporary X-ray security scanners already implement material discrimination via dual-energy multiple view X-ray imagery to enable threat material detection [18]. This use of dual-energy X-ray gives rise to the false-colour mapped appearance of X-ray security imagery (e.g., metals, alloy or hard plastic are shown in blue while less dense objects are shown in green/orange - see Figure 1).

Convolutional Neural Network (CNN) based methods have proven effective in detecting a wide range of object classes within this context [8, 10, 21, 26]. However, the performance of such object detection approaches is heavily reliant on the availability of a substantial volume of labelled X-ray imagery. Unfortunately, the availability of such X-ray imagery datasets suitable for training CNN architectures is limited and also restricted in size and item coverage (e.g. GDXray [14], SIXray [17]).

Commonly, it is challenging to collect sufficient X-ray imagery containing example of prohibited items with large variations in pose, scale and item construction. To overcome this challenge, contemporary data augmentation schemes such as image translation, rotation, flipping and re-scaling are applied to enlarge the availability of otherwise limited training datasets [10]. However, such methods suffer from the fact that the resulting augmented dataset still lacks diversity in terms of prohibited item variation and inter-occlusion emplacement within complex and cluttered X-ray security imagery. This motivates the use of synthetically composed imagery, where such imagery readily enables the introduction of more variability in pose, scale and prohibited item usage in an efficient and readily available way.

In this work, we devise a Synthetically Composited (SC) data augmentation approach via the use of Threat Image Projection (TIP). TIP is an established process within operational aviation security for the monitoring of human operators which uses a smaller collection of X-ray imagery comprising of isolated prohibited objects (only), which are subsequently superimposed onto more readily available benign X-ray security imagery. Here this approach additionally facilitates the generation of synthetic, yet realistic prohibited X-ray security imagery for the purpose of CNN training. Our key contributions are the following: (a) the synthesis of high quality prohibited images from benign X-ray imagery using a documented TIP approach and (b) an extended comparative evaluation on how real and synthetically generated X-ray imagery impacts the performance for prohibited object detection and classification using CNN architectures.

2 Related Work

Traditional computer vision methods that rely on handcrafted features have been applied to prohibited item detection in X-ray security imagery such as Bag of Visual Words (BoVW) [11, 13, 27] and sparse representations [15]. However, the recent advancement in CNN have drawn more attention to prohibited item detection due to significant performance gains within X-ray security imagery [1, 2, 16]. The works of [1, 16] compare handcrafted features with a BoVW based sparse representation to CNN features. These shows that such deep CNN features achieve superior performance with more than 95% accuracy for prohibited item detection. The study of [1] 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 0.99% true positive, 0.01% false positive and 0.994% accuracy for generalized prohibited item detection [1].

Further work on prohibited item under X-ray security imagery is undertaken by Mery *et al.* [13], where regions of interest detection is performed across multiple views of the object. Subsequently, the candidate region obtained from an earlier segmentation step is then matched based on their similarity. This achieves 94.3% true positive and 5.6% false positive across multiple view X-ray security imagery. The work of [2] examines the relative performance of 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 [21], R-FCN [4], and YOLOv2 [20], achieving a maximal 0.88 and 0.97 mAP over 6-class object detection and 2-class firearm detection problems respectively.

To investigate the generalised applicability of CNN within X-ray security imagery, large X-ray imagery datasets are required. Existing public domain datasets such as GDXray [14] contains three major categories of prohibited items, {*Guns, Shurikens, Razor blades*}. However, images in GDXray are provided with lesser clutter and overlap making object detection less challenging than in typical operational conditions. By contrast, the SIXray dataset [17] contains six classes, {*Guns, Knives, Wrenches, Pliers, Scissors, Hammers*}, from cluttered operational imagery. This provides more inter-occluding imagery examples but at the same time provides significantly fewer prohibited item than benign samples akin to an operational (real-world) scenario, where the presence of prohibited items is low within stream-of-commerce (largely benign) X-ray security imagery.

To overcome the limited dataset availability, data augmentation has been used to increase overall dataset diversity. Whilst simple image data augmentation strategies such as translations, flipping and scaling do increase geometric diversity of the imagery they do not increase the appearance or content diversity of the dataset itself [5]. The work of [30] alternatively attempts data augmentation based on an Generative Adversarial Network (GAN) approach but generates synthetic prohibited items in isolation rather than within a full cluttered X-ray security image. By contrast, the work of [9] utilises an approach, similar to the concept of TIP, whereby a prohibited item is superimposed into X-ray security imagery. Therefore, in this work, we explore the feasibility of TIP as a data augmentation strategy to support performance enhancement and evaluation of contemporary deep CNN architectures within the context of prohibited item detection in security X-ray imagery.

3 Proposed Approach

We investigate the use of a full TIP pipeline, based on prior work in the field [19, 23, 24], to generate a range of appearance and contents based dataset variation (Section 3.1). Subsequently, CNN object detection architecture is used to evaluate the TIP based data augmentation approach and compare the performance with real X-ray security imagery (Section 3.2)

3.1 Synthetic X-ray Security Imagery via TIP

Our TIP pipeline consists of three components: *threat image transformation*, *insertion position determination* and *image compositing* as illustrated in Figure 2.

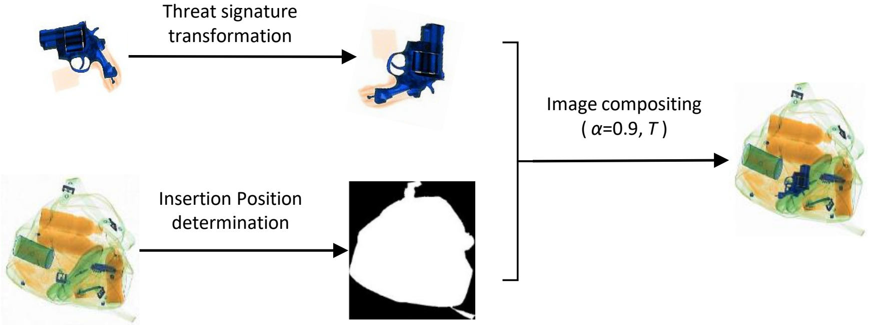
Threat signature transformation
Insertion Position determination
Image compositing
(α=0.9, T)

Figure 2: Threat image projection (TIP) pipeline for synthetically composited image generation.

We use threat (prohibited item) images containing clean, isolated object signatures which can be easily segmented from their plain background via simple thresholding. To diversify the resultant synthetic images, we apply *threat image transformation* via rotating the threat signature by a random angle θ. Although other threat image transformation strategies (e.g., noise, illumination, magnification, etc.) have been explored in [22], our work focuses on the pure combination of our segmented threat signature and a benign X-ray security image, isolating the effects of other data augmentation techniques. We denote this transformed threat image as I_s and the i-th row, j-th column pixel as $I_s(i,j)$.

A valid insertion position within the bag image is determined based on the bag region and the shape of threat signature. Given a bag image I_t, we use morphological operations to extract the bag region. Specifically, the original bag image is firstly binarised by thresholding (Figure 3b) to extract the foreground (target) region for insertion. Due to noise, a simple thresholding process cannot ideally separate background and foreground. We sequentially apply a series of appropriately parameterised morphological operations including dilation (Figure 3c), hole filling (Figure 3d) and erosion (Figure 3e) to identify the largest connected image region as the target for insertion (see Figure 3f). Obviously, a valid insertion of the threat signature has to guarantee the threat signature is completely located inside this target region. To this end, we use a loop to generate a random insertion position until it is a valid one. The selected valid insertion position can be denoted by a binary mask matrix M of the same size as the target baggage image with elements of ones indicating the insertion region.

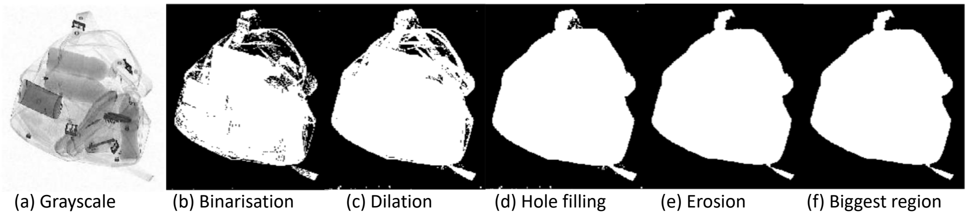

Figure 3: Image segmentation using morphological operations for insertion position determination.

Finally, a threat signature I_s is superimposed onto the target bag image I_t in the selected valid position (denoted by M) to generate a synthetically composited image I_{TIP}. To ensure the plausibility of the composited TIP image, we consider two factors in image blend. Parameter α controls the transparency of the source image I_s ($\alpha = 0.9$). The other parameter is the *threat threshold* T ensuring the consistency of source image with the target image in terms of image contrast. The use of *threat threshold* T aims to remove the high-value pixels of the threat signature so that the inserted threat signature is not visually too bright comparing against the target region where it is superimposed. To calculate the value of T, we first transform the target image I_t to a greyscale image G_t. The threat threshold T can be empirically calculated by:

$$T = min(\exp(\hat{g}^5) - 0.5, 0.95) \tag{1}$$

where $\hat{g}$ is the normalised average intensity of the insertion region within G_t calculated as:

$$\hat{g} = \frac{\sum_{i,j} G_t(i,j) * M(i,j)}{\sum_{i,j} 255 * M(i,j)} \in [0,1] \tag{2}$$

The image compositing can be formulated as follows:

$$I_{TIP}(i,j) = \begin{cases} (1-\alpha)I_t(i,j) + \alpha I_s(i',j'), & M(i,j) = 1 \quad and \quad I_s(i',j') < T * 255, \\ I_t(i,j), & otherwise \end{cases} \tag{3}$$

where $I(i,j)$ denotes the value of pixel in i-th row and j-th column of the image I; $I_s(i',j')$ denotes the pixel in source image corresponding to the pixel of $I(i,j)$. Since the value of T computed by Eq.(1) is in the range of 0.5-0.95, any pixel of the higher value than $T * 255$ in the source image will be ignored during image compositing process.

The proposed TIP approach is able to generate a large number of diverse synthetic X-ray baggage images containing prohibited items whose locations are accessible without any extra cost for training a supervised learning detection model.

3.2 Detection Strategies

We use two representative CNN object detection model, Faster R-CNN [21] and RetinaNet [12], for the purposes of our evaluation.

Faster R-CNN [21] is an object detection algorithm which is the combination of its predecessor Fast R-CNN [6] and Region Proposal Network (RPN). Unlike Fast R-CNN [6], which utilises external region proposal, this architecture has its own region proposal network, which is consists of convolutional layers that generate object proposals and two fully connected layers that predict coordinates of bounding boxes. The corresponding locations and bounding boxes are then fed into objectness classification and bounding box regression layers. Finally the objectness classification layer classify whether a given region proposal is an object or a background region while a bounding box regression layer predicts object localisation, at the end of the overall detection process.

RetinaNet [12] is an object detector where the key idea is to solve the extreme class imbalance between foreground and background classes. To improve the performance, RetinaNet employs a novel loss function called Focal Loss, where it modifies the cross-entropy loss such that it down-weights the loss in easy negative samples so that the loss is focusing on the sparse set of hard samples. Unlike Faster R-CNN [21] which apply two-stage approach, RetinaNet only apply one-stage approach, potentially to be faster and simpler.

4 Experimental Setup

Our experimental setup comprises of real X-ray security imagery dataset and one constructed using the TIP based synthetic compositing approach outlined in Section 3.1. These are evaluated within a common CNN training environment using the CNN architecture outlined in Section 3.2.

Dbf3$_{Real}$ dataset: The Durham Dataset Full Three-class *(Dbf3)* images are generated using a Smith Detection dual-energy X-ray scanner (Figure. 1). It consists of total 7,603 images, which is divided into three classes of prohibited item. In this experiment we uses subsets of the datasets, which consists of three types metallic prohibited item, {*Firearm, Firearm Parts, Knives*}. Out of these three classes, we incorporate 3,192 images of firearms, 1,204 images of firearms parts, and 3,207 images of knives, within cluttered and complex X-ray security dataset.

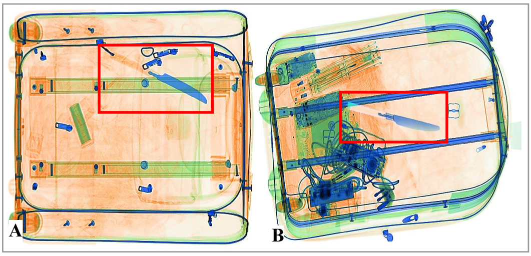

Figure 4: Visual comparison of real (A) and SC (B) X-ray security imagery of prohibited items.

Dbf3$_{SC}$ dataset: The Synthetically Composited (SC) dataset is generated using TIP approach of Section 3.1. We use 3,366 benign X-ray security images, generated by a Smith Detection X-ray scanner, and 123 individual prohibited objects of three classes {*Firearm, Firearm Parts, Knives*}. The prohibited item are composed into the benign images to create synthetically composited X-ray security imagery dataset. We use the same number of images as *Dbf3$_{Real}$* in the synthetically composited dataset. Exemplar images from *Dbf3$_{Real}$* (Figure. 4A) and *Dbf3$_{SC}$* (Figure 4B) are visually realistic and challenging to distinguish from the real images.

Dbf3$_{Real+SC}$ dataset: A subset of *Dbf3$_{Real}$* and subset of *Dbf3$_{SC}$* images are combined to create this dataset, where the numbers of synthetic and real images are are used in equal number to present a data set with 50% of each which is itself the same size as *Dbf3$_{Real}$*.

The CNN architecture (Section 3.2) are trained on a GTX 1080Ti GPU, optimised by Stochastic Gradient Descent (SGD) with a weight decay of 0.0001, learning rate of 0.01 and termination at maximum of 180k epochs. ResNet$_{50}$ and ResNet$_{101}$ are chosen as network backbone to operate within the detection framework of [7]. We split each dataset into training (60%), validation (20%) and test sets (20%) so that each split has similar class distribution. All CNN architecture are initialised with ImageNet [10] pre-trained weights for their respective model [29].

5 Evaluation

Our evaluation considers the comparative performance of CNN architecture to detect prohibited items using real X-ray imagery against prohibited items under synthetic X-ray imagery. We consider mean Average Precision (mAP) as our evaluation criteria following [2].

5.1 Prohibited Item Detection Results

In the first set of experiments (Table 1, upper), prohibited items in X-ray security imagery are detected using the CNN architectures set out in Section 3.2. We use the *Dbf3* dataset consisting of three types of prohibited items $\{$*Firearm, Firearm Parts, Knives*$\}$. To provide performance benchmark, our CNN architectures are firstly trained and evaluated on real images of *Dbf3* ($Dbf3_{Real} \Rightarrow Dbf3_{Real}$). The AP/mAP highlighted in Table 1(upper) denotes the maximal performance achieved. Table 1 shows statistical results of prohibited item detection for Faster R-CNN [21] and RetinaNet [12] architecture using ResNet$_{50}$ and ResNet$_{101}$. Inline with the overall complexity of the network, we observe maximal mAP performance from ResNet$_{101}$ for all three prohibited item classes. In this performance benchmark, we observe that the best performance (mAP = 0.88) is achieved on $Dbf3_{real}$ by Faster R-CNN with ResNet$_{101}$ configuration, as presented in Table 1 (upper).

Train ⇒ Evaluation	Model	Network	Average precision			mAP
			Firearm	Firearm Parts	Knives	
$Dbf3_{Real} \Rightarrow Dbf3_{Real}$	Faster R-CNN [21]	ResNet$_{50}$	0.87	0.84	0.76	0.82
		ResNet$_{101}$	**0.91**	**0.88**	**0.85**	**0.88**
	RetinaNet [12]	ResNet$_{50}$	0.88	0.86	0.73	0.82
		ResNet$_{101}$	0.89	0.86	0.73	0.83
$Dbf3_{SC} \Rightarrow Dbf3_{Real}$	Faster R-CNN [21]	ResNet$_{50}$	0.82	0.77	0.55	0.71
		ResNet$_{101}$	**0.86**	**0.80**	**0.66**	**0.78**
	RetinaNet [12]	ResNet$_{50}$	0.84	0.77	0.53	0.71
		ResNet$_{101}$	0.84	0.76	0.54	0.72
$Dbf3_{Real+SC} \Rightarrow Dbf3_{Real}$	Faster R-CNN [21]	ResNet$_{50}$	0.85	0.79	0.65	0.76
		ResNet$_{101}$	**0.87**	**0.81**	**0.74**	**0.81**
	RetinaNet [12]	ResNet$_{50}$	0.85	0.81	0.64	0.76
		ResNet$_{101}$	0.86	0.80	0.63	0.76

Table 1: Detection results of varying CNN architecture trained on: Upper $\rightarrow$ $Dbf3_{Real}$, Middle $\rightarrow$ $Dbf3_{SC}$ and Lower $\rightarrow$ $Dbf3_{Real+SC}$. All models are evaluated on set of real X-ray security imagery.

In second set of experiments (Table 1, middle), the CNN architecture are trained on the synthetic X-ray imagery ($Dbf3_{SC}$) achieve 0.78 mAP when tested on same set of real X-ray imagery ($Dbf3_{Real}$) of Table 1 (upper). Even though the performance is lesser when compared with former results (Table 1, upper), this experimental setting does not require any manual image labelling (as TIP insertion positions are known) and yet achieves surprisingly good performance on a standard benchmark. The performance gap between CNN architecture trained on real and synthetically composited X-ray imagery is attributable to the domain shift problem whereby the distribution of training and test data differ. In the first experiment (Table 1, upper), the training and test data are from the same distribution since they are created by randomly dividing data captured under the same experimental conditions. By contrast, in this second experimental setup (Table 1, middle), the prohibited items used for the synthetic X-ray imagery ($Dbf3_{SC}$) data are different from those in the test X-ray imagery ($Dbf3_{Real}$) data. It is also noteworthy that prohibited images used for generating synthetic X-ray imagery ($Dbf3_{SC}$) data is a smaller set of prohibited item instances than in the real training images. As a result, CNN architecture trained on synthetic data have larger generalisation errors than those trained on real data. However, when tested on synthetic X-ray imagery ($Dbf3_{SC}$) data (Table 2), however, CNN architecture trained with real or synthetic CNN architecture have comparable performance. These experimental results show that it is essential to have diverse prohibited item signatures in the training data to improve the generalisation. It also largely explains why overall performance in Table 2 (showing evaluation on the synthetic dataset, $Dbf3_{SC}$) is significantly higher than overall performance in Table 1 (evaluation is on the real dataset, $Dbf3_{Real}$).

Train ⇒ Evaluation	Model	Network	Average precision			mAP
			Firearm	Firearm Parts	Knives	
$Dbf3_{Real}$ ⇒ $Dbf3_{SC}$	Faster R-CNN [21]	ResNet$_{50}$	0.88	0.87	0.84	0.87
		ResNet$_{101}$	**0.92**	**0.92**	**0.89**	**0.91**
	RetinaNet [12]	ResNet$_{50}$	0.89	0.87	0.83	0.86
		ResNet$_{101}$	0.90	0.88	0.85	0.88
$Dbf3_{SC}$ ⇒ $Dbf3_{SC}$	Faster R-CNN [21]	ResNet$_{50}$	0.90	0.88	0.83	0.87
		ResNet$_{101}$	**0.93**	**0.92**	**0.86**	**0.91**
	RetinaNet [12]	ResNet$_{50}$	0.91	0.89	0.84	0.88
		ResNet$_{101}$	0.91	0.89	0.83	0.86
$Dbf3_{Real+SC}$ ⇒ $Dbf3_{SC}$	Faster R-CNN [21]	ResNet$_{50}$	0.89	0. 86	0.83	0.86
		ResNet$_{101}$	**0.91**	**0.89**	**0.87**	**0.89**
	RetinaNet [12]	ResNet$_{50}$	0.90	0.87	0.83	0.87
		ResNet$_{101}$	0.90	0.88	0.84	0.87

Table 2: Detection results of different CNN architecture trained on: Upper → $Dbf3_{Real}$, Middle → $Dbf3_{SC}$ and Lower → $Dbf3_{Real+SC}$. All models are evaluated on set of SC dataset.

In the third set of experiments (Table 1, bottom), we evaluate the effectiveness of synthetic X-ray imagery by combining it with real images of *Dbf3* to create $Dbf3_{Real+SC}$ dataset, as explained in the Section 4. We evaluate the testing sets of images from real *Dbf3* (Table 1) and synthetically composite (Table 2) datasets. Surprisingly, the combination of real and synthetic imagery data does not improve the results (e.g. 0.81 vs 0.88 on $Dbf3_{Real}$ and 0.89 vs 0.91 on $Dbf3_{SC}$ with Faster R-CNN and ResNet101). This can also be explained by the domain shift problem mentioned previously. Possibly this data combination can perform well if we apply domain adaptation techniques [28] explicitly. In addition, we may also need to evaluate the quality of the TIP solution that underpins our work further.

5.2 Qualitative Examples

Exemplar prohibited items detection results from Faster R-CNN [21] with ResNet$_{101}$ are depicted in Figure 5, using real (top row) and synthetic (bottom row) training imagery. These results illustrate that the synthetically composited imagery using TIP techniques can be effective in training detection architectures for prohibited item detection in cluttered X-ray security imagery.

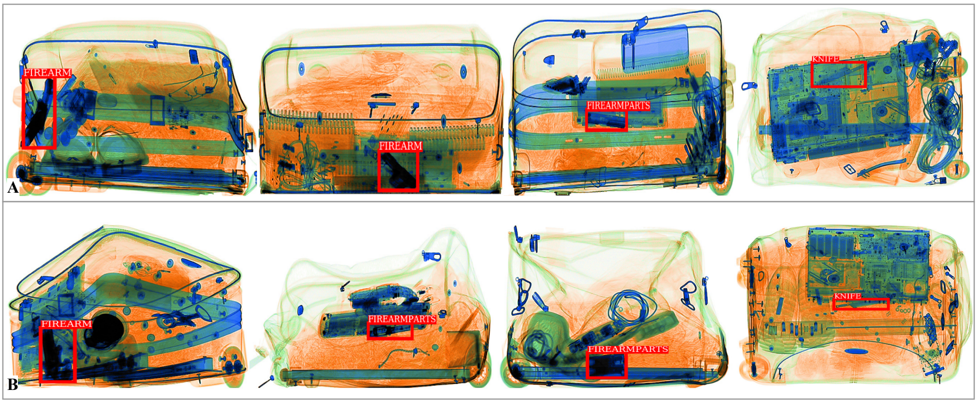

Figure 5: Exemplar detection of prohibited items in red box using Faster R-CNN [21] and trained on (A) *Dbf3*$_{Real}$ and (B) *Dbf3*$_{SC}$ images.

We also visually inspect the detection results to investigate the performance difference when training the models using real and synthetic data. By comparing the results depicted in Figures 6A1 and 6B1, the model trained with synthetic data fails to detect the knives since such type of knives have very different appearance from the ones we used to generate the synthetic imagery. On the other hand, from Figures 6A2 and 6B2 we can see that the model trained on synthetic imagery has mistakenly detected something benign as a knife. These results account for the low performance for knife detection observed in Table 1. As a result, we need to either use more diverse threat signatures for data synthesis or particular domain adaptation techniques to tackle the potential domain shift problem identified previously.

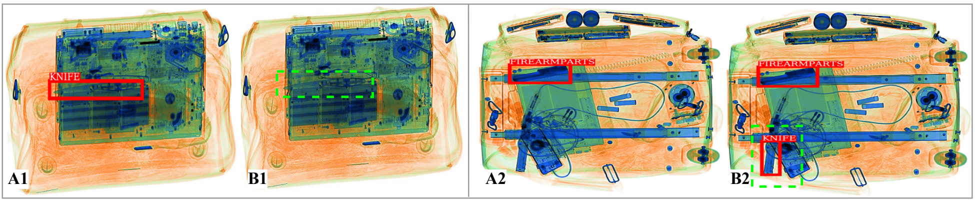

Figure 6: Exemplar prohibited item detection (by Faster R-CNN [21]) using *Dbf3*$_{Real}$ (A1,A2) and *Dbf3*$_{SC}$ (B1,B2) training datasets. Green dashed box in B1 fails to detect, where in B2 wrongly detects as *knife*.

6 Conclusion

This work explores the possibility of generating synthetically composite X-ray security imagery for training of CNN architecture to bypass the collecting a large amount of hand-annotated real-world X-ray baggage imagery. We synthesise high-quality synthetically composited X-ray images using TIP approach and we present an extensive comparison on how real and synthetic X-ray security imagery affects the performance of CNN architecture for prohibited object detection in cluttered X-ray baggage images. Our experimental comparison demonstrates Faster R-CNN achieves the highest performance with mAP: 0.88 when trained on *Real* data (the good), followed by *Real+Synthetic* (the bad) and *Synthetic* (the ugly) over a three-class, {*Firearms, Firearm parts, Knives*}, prohibited item detection problem. This demonstrates a strong insight into the benefits of using real X-ray training data, also challenge and promise of using synthetic X-ray imagery.

In our future work, it is worth further investigating how to improve the effectiveness of synthetically composited imagery for training CNN architecture. Based on other work [30], a potential direction is to generate more diverse prohibited items images using generative adversarial networks (GAN). The generated prohibited item images then could be used for generating synthetic baggage images using TIP or similar.

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 *IEEE International Conference on Image Processing*, pages 1057–1061. IEEE, 2016.

[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*, 13(9):2203–2215, 2018.

[3] G. Blalock, V. Kadiyali, and D. H Simon. The impact of post-9/11 airport security measures on the demand for air travel. *The Journal of Law and Economics*, 50(4): 731–755, 2007.

[4] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 379–387. NIPS, 2016.

[5] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *International Symposium on Biomedical Imaging*, pages 289–293. IEEE, 2018.

[6] R Girshick. Fast R-CNN. In *International Conference on Computer Vision*, pages 1440–1448. IEEE, 2015.

[7] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. Detectron. https://github.com/facebookresearch/detectron, 2018.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*. IEEE, June 2016.

[9] D.K. Jain et al. An evaluation of deep learning based object detection strategies for threat object detection in baggage security imagery. *Pattern Recognition Letters*, 120: 112–119, 2019.

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

[11] M. E. Kundegorski, S. Akçay, 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*, pages 1–6. IEEE, 2016.

[12] T-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *International Conference on Computer Vision*, pages 2980–2988. IEEE, 2017.

[13] D. Mery, V. Riffo, I. Zuccar, and C. Pieringer. Automated X-ray object recognition using an efficient search algorithm in multiple views. In *Conference on Computer Vision and Pattern Recognition workshops*, pages 368–374. IEEE, 2013.

[14] D. Mery, V. Riffo, U. Zscherpel, G. Mondragón, I. Lillo, I. Zuccar, H. Lobel, and M. Carrasco. Gdxray: The database of X-ray images for nondestructive testing. *Journal of Nondestructive Evaluation*, 34(4):42, 2015.

[15] D. Mery, E. Svec, and M. Arias. Object recognition in baggage inspection using adaptive sparse representations of X-ray images. In *Image and Video Technology*, pages 709–720. Springer, 2015.

[16] 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. *Transactions on Systems, Man, and Cybernetics: Systems*, 47(4):682–692, 2016.

[17] 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 *Conference on Computer Vision and Pattern Recognition*, pages 2119–2128. IEEE, 2019.

[18] A. Mouton and T. P. Breckon. A review of automated image understanding within 3D baggage computed tomography security screening. *Journal of X-ray Science and Technology*, 23(5):531–555, 2015.

[19] E. C. Neiderman and J. L. Fobes. Threat image projection system, 2005. US Patent 6,899,540.

[20] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Conference on Computer Vision and Pattern Recognition*, pages 7263–7271. IEEE, 2017.

[21] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[22] T. W. Rogers, N. Jaccard, E. D. Protonotarios, J. Ollier, E. J. Morton, and L. D. Griffin. Threat image projection (TIP) into X-ray images of cargo containers for training humans and machines. In *International Carnahan Conference on Security Technology*, pages 1–7. IEEE, 2016.

[23] A. Schwaninger, D. Hardmeier, and F. Hofer. Measuring visual abilities and visual knowledge of aviation security screeners. In *International Carnahan Conference on Security Technology*, pages 258–264. IEEE, 2004.

[24] A. Schwaninger, S. Michel, and A. Bolfing. A statistical approach for image difficulty estimation in X-ray screening using image measurements. In *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization*, pages 123–130. ACM, 2007.

[25] A. Schwaninger, A. Bolfing, T. Halbherr, S. Helman, A. Belyavin, and L. Hay. The impact of image based factors and training on threat detection performance in X-ray screening. In *Conference on Research in Air Transportation*, pages 317–324, 2008.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017.

[27] D. Turcsany, A. Mouton, and T. P. Breckon. Improving feature-based object recognition for X-ray baggage security screening using primed visualwords. In *International Conference on Industrial Technology*, pages 1140–1145. IEEE, 2013.

[28] Q. Wang, P. Bu, and T.P. Breckon. Unifying unsupervised domain adaptation and zero-shot visual recognition. In *International Joint Conference on Neural Networks*, 2019.

[29] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[30] J. Yang, Z. Zhao, H. Zhang, and Y. Shi. Data augmentation for X-ray prohibited item images using generative adversarial networks. *IEEE Access*, 7:28894–28902, 2019.