On the Evaluation of Prohibited Item Classification and Detection in Volumetric 3D Computed Tomography Baggage Security Screening Imagery

Qian Wang
Department of Computer Science
Durham University
Durham, UK

Neelanjan Bhowmik
Department of Computer Science
Durham University
Durham, UK

Toby P. Breckon
Department of {Computer Science | Engineering}
Durham University
Durham, UK

***Abstract*—X-ray Computed Tomography (CT) based 3D imaging is widely used in airports for aviation security screening whilst prior work on prohibited item detection focuses primarily on 2D X-ray imagery. In this paper, we aim to evaluate the possibility of extending the automatic prohibited item detection from 2D X-ray imagery to volumetric 3D CT baggage security screening imagery. To these ends, we take advantage of 3D Convolutional Neural Networks (CNN) and popular object detection frameworks such as RetinaNet and Faster R-CNN in our work. As the first attempt to use 3D CNN for volumetric 3D CT baggage security screening, we first evaluate different CNN architectures on the classification of isolated prohibited item volumes and compare against traditional methods which use hand-crafted features. Subsequently, we evaluate object detection performance of different architectures on volumetric 3D CT baggage images. The results of our experiments on *Bottle* and *Handgun* datasets demonstrate that 3D CNN models can achieve comparable performance ($\sim$98% true positive rate and $\sim$1.5% false positive rate) to traditional methods but require significantly less time for inference (0.014s per volume). Furthermore, the extended 3D object detection models achieve promising performance in detecting prohibited items within volumetric 3D CT baggage imagery with $\sim$76% mAP for bottles and $\sim$88% mAP for handguns, which shows both the challenge and promise of such threat detection within 3D CT X-ray security imagery.**

***Index Terms*—3D volumetric data, deep convolutional neural network, X-ray computed tomography, baggage data, classification, object detection.**

I. Introduction

X-ray baggage security screening is widely used to maintain aviation security. Currently, multi-view X-ray is predominantly used in aviation security for cabin baggage screening. This traditional baggage screening process, using 2D X-ray scanners, has the disadvantage of both inter-object occlusion and clutter within any given image projection of the scanned baggage item. As a result, it poses a considerably challenging visual search task for the human operators to discover the prohibited items (e.g., liquids, firearms, knives, etc.) overlapped with other benign items (e.g., electronic devices) within a constrained time frame. For this reason, passengers are currently required to divest large electronic devices and liquids which decreases checkpoint throughput significantly. Furthermore, human operator performance can be subjective and is heavily affected by many factors such as the experience, fatigue, monotony and concentration, although many successful measures have been taken to alleviate the problem in practice (e.g., Threat Image Projection (TIP) [1], [2] and shorter shift rotations [3]).

By leveraging recent advances in object classification and detection, significant progress has been made in automatic prohibited item detection within 2D X-ray imagery [4]. The use of deep learning techniques allows real-time and accurate detection of prohibited items even in cluttered X-ray images [5]–[7]. However, performance can be affected when the baggage contains significant clutter and inter-object occlusion due to the fundamental limitation of projected 2D X-ray imagery. To improve the detection rate without affecting the checkpoint throughput, airports are currently increasing the use of 3D CT screening which does not require the removal of electronic devices and liquids during baggage screening. The reconstructed 3D CT images provide more information and make it possible for the human operators to inspect the 3D CT images from differing views. However, current technology does not facilitate the automatic detection of (non-explosive) prohibited items such as weapons and liquid containers. It is unknown *if the success of deep learning approaches in 2D X-ray imagery can be similarly replicated in volumetric 3D CT imagery for baggage security screening* and *whether the 3D CNN based approaches are efficient enough for operational viability?*

To answer the above questions, in this paper we extend the prohibited item classification and detection methods from 2D to 3D imagery and evaluate their effectiveness in real volumetric 3D CT baggage security screening imagery. Firstly, we look into the task of 3D object classification for isolated prohibited items in volumetric 3D CT data. We investigate different CNN architectures including ResNet [8] with variable depths and Voxception-ResNet [9]. We also evaluate the effectiveness of data and feature augmentation techniques in 3D CNN based classification. As for the detection problem, we consider two successful object detection frameworks for 2D imagery: Faster R-CNN [10] and RetinaNet [11].

The contributions of this work are summarized as follows:

– the first attempt to use deep CNN models for the prohibited item classification and detection within volumetric 3D CT baggage imagery to our best knowledge;
– an evaluation of different 3D CNN models in the classification of prohibited items within volumetric 3D CT baggage imagery and the effect of data/feature augmentation;
– an evaluation of prohibited item detection within volumetric 3D CT baggage imagery using 3D Faster R-CNN and 3D RetinaNet CNN architectures.

II. Related Work

The work presented in this paper is closely related to some prior art in two aspects which we briefly discuss in this section: *3D baggage imagery analysis* and *3D CNN*.

A. 3D Baggage Imagery Analysis

To enable automatic baggage screening using 3D CT imagery, a variety of studies have been carried out in recent years [2], [12]–[19].

One research direction is object segmentation based on the material and morphological structure [12], [17], [19]. Specifically, Mouton et al. [17] proposed a two-stage approach for object segmentation within 3D CT imagery. A CT volume is firstly coarsely segmented based on the voxel intensity ranges of pre-defined materials. Subsequently, a variety of shape descriptors are computed as features for the random forest classifier to determine a segment resulted from the first stage is good (containing only one object) or bad (containing multiple objects and hence need further segmentation). Wang et al. [19] studied the issue of object segmentation and classification in 3D CT imagery and focused mainly on the material characteristics without considering any specific prohibited item (e.g., firearm, knife, etc.). An approach to 3D segmentation was proposed based on recursive morphological operations and the Support Vector Machines (SVM) were employed for the classification of three types of materials.

3D object classification was studied in [13], [14], [16] where a binary classifier was formulated to distinguish the objects of interest (i.e. handgun or bottle) from the background volumes which contain cluttered content (e.g., books, clothes and etc.). The bag-of-word features and a SVM classifier were used in these studies (denoted Cortex [13], Codebook [16] and ERC [14] in Table III). Isolated 3D CT volumes of prohibited items are manually cropped from the baggage CT images to form the positive sample set which are also employed here in our work for the evaluation of 3D CNN based classification.

More recently, Wang et al. [2] present an approach to 3D threat image projection which can be used to generate realistic and plausible volumetric 3D CT baggage images with superimposed threat object signatures. This technique can be used for training not only human operators for baggage security screening but also machine learning algorithms for automatic prohibited item detection without time-consuming data collection and manual annotation such as in [1].

B. 3D Convolutional Neural Networks

3D CNN models are widely used for object classification and detection within varying data modalities such as LiDAR point cloud [20], [21], RGB-Depth data [22], 3D Computer Aided Design (CAD) models [9] and medical CT imagery [23], [24].

VoxelNet [21] is an end-to-end 3D object detector specially designed for LiDAR data. It consists of three modules: feature learning network (subdivide the point cloud into many subvolumes/voxels, feature engineering + fully connected neural network), convolutional middle layer (3D convolution applied to the stacked voxel feature volumes, each subvolume/voxel is a feature vector) and region proposal networks. VoxNet [20] in a more generic model being able to handle different types of 3D data including LiDAR point cloud, CAD and RGBD data. Qi et al. [25] improved the performance of VoxNet by introducing the auxiliary subvolume supervision to alleviate the overfitting issue.

RGB-Depth data can also be processed using 3D CNN by firstly extracting proposals from 2D RGB images using a 2D object detector and transforming the proposals and corresponding depth information into 3D point clouds [22]. The generated 3D point clouds can be further explored by 3D CNN models such as PointNet [26].

These models designed for point clouds, RGBD data, CAD models or medical CT images are not readily transferable to our volumetric 3D CT imagery for baggage security screening since the modality of input data for 3D CNN can differ significantly. However, the design of 3D CNN architectures and the training strategies used in existing work can be repurposed towards our prohibited item classification and detection within 3D CT baggage imagery.

III. Method

In this section, we describe the methods used in our work for prohibited item classification and detection within volumetric 3D CT baggage imagery. We firstly consider a classification problem to evaluate the possibility of discriminating the isolated prohibited item signatures from benign CT volumes. Subsequently, we consider the more realistic detection problem which aims to not only classify the prohibited item within a baggage CT image but also localises it by generating a 3D bounding box around the target object.

A. 3D Prohibited Item Classification

Our prohibited item classification is formulated as a binary classification problem in this work to evaluate the effectiveness of 3D CNN models. Specifically, given a CT volume as the input, the 3D CNN model aims to determine if the volume contains a prohibited item signature (positive sample) or not (negative sample). The positive samples are manually cropped from baggage volumes, hence they can have varying dimensions and orientations. As a result, we need to preprocess the input samples to a common voxel scaling before feeding them into the 3D CNN.

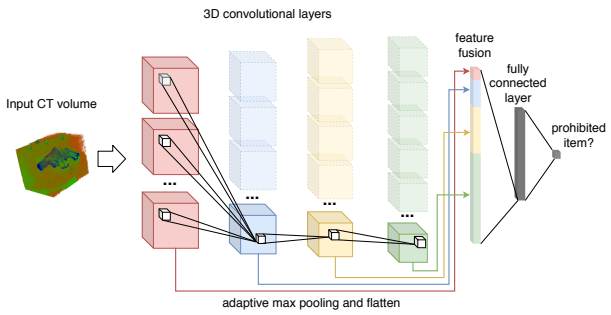

Fig. 1. Our 3D CNN architecture for classification with rich feature fusion based on ResNet [8] (four CNN layers used for feature fusion are depicted here whilst many other intermediate layers are omitted; the output layer has one node for our binary classification problems).

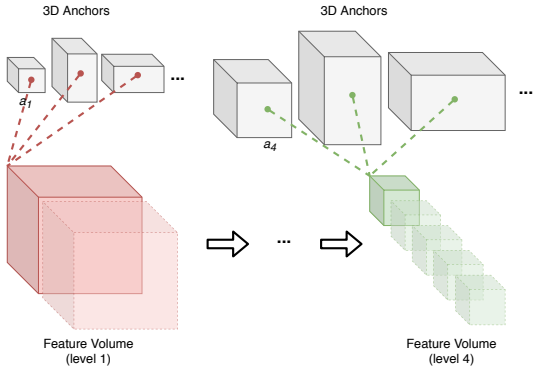

Fig. 2. An illustration of 3D anchors built on multi-level feature volumes (small anchors are more densely built with references to all voxels within a lower-level 3D feature volume).

1) 3D CNN Model: We consider ResNet [8] and Voxception-ResNet (VRN) [9] in our evaluation. ResNet was extended to the 3D version by Chen et al. [27] for medical CT image analysis. To address the issue of variable sizes of prohibited items, we employ the idea of rich features [28] to explore the multi-scale feature volumes. Specifically, we augment the features for fully-connected layers by fusing the feature volumes generated by multiple intermediate convolutional layers. The proposed architecture of *rich feature* ResNet is illustrated in Figure 1. The architecture is composed of four sequential blocks, each of which contains multiple 3D convolutional layers. By stacking different numbers of layers, we investigate variants of ResNet (ResNet_{10}, ResNet_{18} and ResNet_{34}) in our experiments. Even deeper ResNet models (e.g., ResNet_{50} and ResNet_{101}) are also investigated. However, these deeper models suffer with convergence problems.

Alternatively, we also consider a variation of deeper ResNet: Voxception-ResNet (VRN). VRN is designed by combining the ideas of Inception-style [29] networks and ResNet in a 3D CNN framework. It takes advantage of the Inception-style architectures for multi-scale visual information exploration and the advantage of residual connections for efficient training.

2) Data Pre-processing and Augmentation: Correctly designed data pre-processing and augmentation strategies are beneficial for training CNN models for small datasets. Here we consider two strategies for data pre-processing and augmentation: *rescaling* and *rotation*. Since our chosen CNN architecture uses an adaptive pooling layer before the fully-connected layers to handle the variable dimensions of the feature volumes caused by the varying input sizes, the input volumes are not required to have the same size. The *rescaling* aims to restrict three dimensions (i.e. height, width and depth) of input volumes within a limited range. Specifically, we rescale the input 3D volumes to have dimensions no greater than a pre-defined number of voxels in any dimension by down-sampling a given volume $\boldsymbol{V} \in \mathbb{R}^{H \times W \times D}$ by factors of $max\{1, \lfloor H/s \rfloor\}$, $max\{1, \lfloor W/s \rfloor\}$ and $max\{1, \lfloor D/s \rfloor\}$ for all three dimensions respectively. The hyper-parameter scaling value s is empirically chosen as 32 for favourable classification performance.

During training, the 3D input volumes are randomly rotated in three planes (i.e., xy, yz and xz) with a fixed probability to augment the training data. The augmentation of *rotation* is enabled randomly in one of the three planes and the rotation angles are restricted to $\{90, 180, 270\}$ degrees. Without this restriction the volumes after rotation become more complicated by the requirement for resampling and zero padding and this poses an additional challenge for training.

B. 3D Prohibited Item Detection

Prohibited item detection within volumetric 3D CT baggage imagery is a more challenging problem aiming to both localise and classify the prohibited items simultaneously. We look into the possibility of extending 2D object detection frameworks to resolve this problem arising from real-world applications of aviation security. Faster R-CNN [10] and RetinaNet [11] are considered for their superior performance in 2D object detection within this domain [1], [5].

Faster R-CNN consists of three modules: Feature Extraction Network, Region Proposal Network and Region of Interest (RoI) pooling. We use ResNet_{50} and ResNet_{101} as the backbone networks for feature extraction. To handle the object scale variability, we use a Feature Pyramid Network similar to the rich feature extraction strategy used in the classification models. By contrast, RetinaNet is a one-stage object detection approach. Again, ResNet_{50} and ResNet_{101} are used as the backbone networks for feature extraction. For both methods the anchors are defined on multiple levels of feature volumes.

The anchor size is an important factor affecting the performance of detectors used in our evaluation. We follow the work on 2D object detection [10], [11] and extend it to our 3D object detection frameworks. We use a set of anchor sizes and ratios to generate diverse 3D anchors in four feature pyramid levels. Specifically, we set anchor sizes as $\{a_l\}, l = 1, 2, 3, 4$ for four feature pyramid levels respectively. The anchor ratios are uniformly set as (*height:width:depth*) $\{1:2:\sqrt{2}; 1:1:1; 2:1:\sqrt{2}\}$. These ratios are empirically selected rather than generated by k-means [30] clustering over training bounding boxes for the fact that the prohibited items within baggage can have arbitrary orientations leading to arbitrary box ratios even

the items themselves have fixed dimension ratios. With the combination of anchor sizes and ratios, three 3D anchors are generated for each voxel in the feature volume for Faster R-CNN. In the RetinaNet framework, the anchors are augmented by adding extra anchor sizes of $\{a_l 2^{1/3}, a_l 2^{2/3}\}$ for all feature pyramid levels $l = 1, 2, 3, 4$ [11]. As a result, for each voxel in a feature volume nine 3D anchors are generated for RetinaNet. The anchor sizes of lower-level feature volumes should be smaller since the anchors are more densely built as shown in Figure 2. The effect of varying anchor sizes will be evaluated in our experiments.

IV. Experimental Setup

In this section, we describe the experimental setup for the evaluation of prohibited item classification and detection within baggage CT volumes. We describe the datasets used in our experiments and implementation details of the classification and detection methods.

A. Dataset

We use the same datasets as employed in [16] for classification and detection tasks. The data was obtained from a CT80-DR dual-energy baggage-CT scanner manufactured by Reveal Imaging Inc. Two object categories (i.e. bottles and handguns) are considered in our experiments for proof-of-concept.

For classification, we use the manually isolated CT volumes from the original baggage CT volumes. These isolated CT volumes form two independent datasets. The *Bottle* dataset contains 1704 isolated CT volumes among which 526 are positive samples (i.e. with bottles) and 1178 are negative samples (i.e. without bottles). The *Handgun* dataset contains 1255 isolated CT volumes among which there are 284 positive samples and 971 negative ones. Some exemplar isolated CT volumes of bottles and handguns are shown in Figure 3. The same ten-fold cross-validation used in [16] was employed in our experiments for classification.

For detection, we use the original whole baggage CT volumes. Again two datasets (i.e. *Bottle* and *Handgun*) are considered independently in our experiments. There are 305 baggage volumes in the *Bottle* dataset and 526 bottle signatures are annotated by 3D bounding boxes. There are 267 baggage volumes in the *Handgun* dataset within which 282 handgun signatures are annotated by 3D bounding boxes. We divide the dataset into training (80%) and test (20%) subsets randomly. Three random splits are generated for each dataset in our experiments.

B. Implementation Detail

The classification and detection models evaluated in this work are implemented in PyTorch [31]. In the classification experiments, we use the Adam [32] optimiser with the learning rates of 0.0001 and 0.00001 for the Bottle and Handgun datasets respectively. In the 3D detection experiments, our models are also optimised by the Adam with a learning rate of $1e-4$ and stop training at 150 iterations. All experiments are conducted on a GTX 1080Ti GPU.

V. Evaluation

We present and evaluate experimental results of both classification and detection in this section.

A. Evaluation Criteria

For the classification task, our model performances are evaluated in terms of True Positive rate (TPR%) and False Positive rate (FPR%). The mean and standard deviations over 10-fold cross validation are reported as the experimental results. For the detection task, we set the IoU (Intersection over Union; of ground truth and predicted 3D bounding boxes) threshold as 0.1 since it is significantly more challenging to get good overlapping bounding boxes in 3D object detection than that in 2D detection where an IoU threshold greater than 0.5 is typically used [5]. Precision (%) and Recall (%) are calculated by thresholding the associated classification score at 0.9. In addition, Average Precision (%) is reported to evaluate the overall model performance. All these evaluation criteria are reported as mean $\pm$ standard deviation over the three random splits of each dataset.

B. 3D Classification

3D ResNet and Voxception-ResNet (VRN) are evaluated for 3D prohibited item classification in our experiments with evaluation results shown in Tables I and II.

We firstly investigate the effect of different data preprocessing and augmentation strategies. For each model, the performance is reported when different combinations of strategies are used. It is obvious that *rescaling* with $s = 32$ is important to good classification performance. Without *rescaling* the TPR is low and the FPR is high for ResNet_{10} and ResNet_{18} models on the *Bottle* dataset and in the rest cases all models can not even converge (denoted by n/a in Tables I-II). The use of *rescaling* significantly improves performance in all situations. Recall that *rescaling* tends to reduce the difference of three dimensions (i.e., height, width and depth) of the input CT volumes, it is crucial to ensure the input volumes to have similar dimensions and suitable sizes so that the information loss can be avoided in the adaptive pooling layer.

The use of *rich features* benefits the ResNet models on both datasets with increased TPR and reduced FPR consistently (Tables I, II). However, rich features do not make a difference for the VRN model. The reason is the VRN model has already employed the Inception-style architecture which is exactly for the fusion of multiple features learned with different kernel sizes. These results demonstrate the fact that rich features characterising information underlying different spatial scales are crucial for good classification performance and it can be implemented in different ways (e.g., multi-level feature volume fusion and inception-style network module).

The effect of *rotation* varies on two datasets. For the *Bottle* dataset, the use of *rotation* does not improve the performance of any model (Table I). However, the data augmentation with *rotation* benefits the classification of handguns significantly, especially for the ResNet models (Table II). One potential

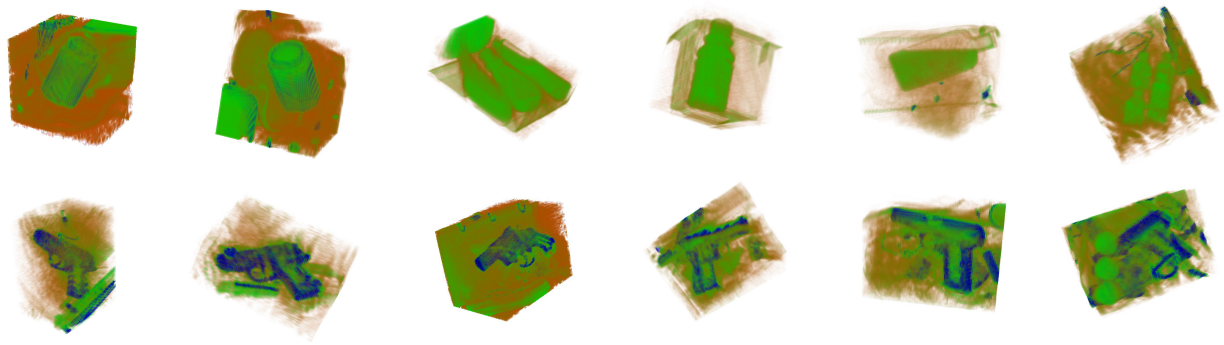

Fig. 3. Exemplar CT volumes of isolated prohibited items (bottles in the upper row and handguns in the bottom row) used in our classification experiments.

explanation is that the rotation operation applied into handgun signatures can generate more diversity than that into bottles.

By comparing different models in Tables I,II, the VRN model achieves the highest TPR (98.9% on bottles and 97.5% on handguns) with moderate FPR (1.4% on bottles and 1.5% on handguns). As for the ResNet models, those having more layers generally perform better than the shallow ones. However, our results of ResNet_{50} and ResNet_{101} (not shown in tables) for classification demonstrate that deeper models are more difficult to converge with a limited number of training samples.

Table III shows the comparison results of classification with prior work [13], [14], [16]. Overall, the best performing 3D CNN model VRN can achieve better results than the earlier Cortex [13] and Codebook [16] methods but is slightly worse than ERC [14] which, however, suffers from heavy computation burdens for computing dense 3D visual descriptors (187s per volume [14]). By contrast, our 3D CNN models are more efficient in the inference stage especially via parallelised GPU computation (0.0142s per volume).

TABLE I
CLASSIFICATION RESULTS OF VARYING 3D CNN ARCHITECTURES ON THE **BOTTLE** DATASET.

Model	Augmentation			TPR (%)	FPR (%)
	Res	Rot	RF		
ResNet_{10}	✗	✗	✗	57.1 ± 16.2	26.7 ± 7.4
	✓	✗	✗	93.3 ± 3.5	1.6 ± 1.0
	✓	✗	✓	95.4 ± 2.9	0.9 ± 0.9
	✓	✓	✓	95.4 ± 4.0	0.9 ± 0.8
ResNet_{18}	✗	✗	✗	61.3 ± 12.8	26.5 ± 22.6
	✓	✗	✗	94.7 ± 4.3	**0.4 ± 0.8**
	✓	✗	✓	94.8 ± 2.4	1.1 ± 2.5
	✓	✓	✓	96.0 ± 3.9	0.8 ± 0.8
ResNet_{34}	✗	✗	✗	n/a	n/a
	✓	✗	✗	93.3 ± 6.7	2.9 ± 5.7
	✓	✗	✓	94.9 ± 3.5	0.7 ± 0.9
	✓	✓	✓	94.9 ± 2.6	0.8 ± 1.1
Voxception-ResNet	✗	✗	✗	86.1 ± 10.4	14.7 ± 11.7
	✓	✗	✗	98.9 ± 1.6	0.6 ± 0.7
	✓	✗	✓	97.7 ± 2.4	0.8 ± 0.7
	✓	✓	✗	**98.9 ± 1.0**	1.4 ± 1.1

TABLE II
CLASSIFICATION RESULTS OF VARYING 3D CNN ARCHITECTURES ON THE **HANDGUN** DATASET.

Model	Augmentation			TPR (%)	FPR (%)
	Res	Rot	RF		
ResNet_{10}	✗	✗	✗	n/a	n/a
	✓	✗	✗	80.5 ± 6.5	10.8 ± 1.1
	✓	✗	✓	83.7 ± 6.0	8.3 ± 2.9
	✓	✓	✓	84.9 ± 9.5	11.3 ± 4.1
ResNet_{18}	✗	✗	✗	n/a	n/a
	✓	✗	✗	82.6 ± 10.9	10.3 ± 1.6
	✓	✗	✓	85.1 ± 8.1	8.8 ± 3.8
	✓	✓	✓	89.4 ± 9.1	9.6 ± 5.2
ResNet_{34}	✗	✗	✗	n/a	n/a
	✓	✗	✗	85.5 ± 6.0	7.6 ± 4.9
	✓	✗	✓	89.7 ± 5.5	**0.8 ± 0.8**
	✓	✓	✓	93.4 ± 4.6	3.6 ± 2.2
Voxception-ResNet	✗	✗	✗	n/a	n/a
	✓	✗	✗	96.1 ± 4.9	1.5 ± 2.1
	✓	✗	✓	94.7 ± 5.4	1.0 ± 1.0
	✓	✓	✗	**97.5 ± 2.4**	1.5 ± 1.1

TABLE III
COMPARISON RESULTS OF CLASSIFICATION WITH PRIOR WORK.

Method	Bottles		Guns	
	TPR (%)	FPR (%)	TPR (%)	FPR (%)
Cortex [13]	96.6 ± 3.2	1.0 ± 1.6	96.8 ± 2.6	1.1 ± 0.9
Codebook [16]	89.3 ± 5.5	3.0 ± 1.4	97.3 ± 3.4	1.8 ± 1.7
ERC [14]	**98.9 ± 0.7**	**0.6 ± 0.3**	**99.7 ± 0.5**	**0.3 ± 0.2**
VRN (Ours)	**98.9 ± 1.0**	1.4 ± 1.1	97.5 ± 2.4	1.5 ± 1.1

C. 3D Object Detection

For prohibited item (i.e., bottles and handguns) detection within 3D X-ray CT baggage screening images, we employ Faster R-CNN [10] and RetinaNet [11] CNN architectures as set out in Section III-B. The detection results using ResNet_{50} and ResNet_{101} backbone netwroks are presented in the Tables IV and V.

In our detection experiments we investigate the effect of different factors, such as *resampling*, *anchor box size*, and *confidence threshold*. We report the best performing combinations by varying configurations for both detection models. The resampling is essential to achieve good detection performance. We observe, for both detection architectures on both datasets, resampling CT volume by $1/3$ in all three dimensions signif-

TABLE IV
DETECTION RESULTS OF VARYING 3D CNN ARCHITECTURES ON THE **BOTTLE** DATASET.

Model	Network	Anchor size	Score threshold=0.9		Average Precision (%)
			Precision (%)	Recall (%)	
Faster R-CNN [10]	ResNet_{50}	4-8-16-32	80.41 ± 3.77	62.47 ± 5.47	58.84 ± 5.91
		6-12-24-48	85.99 ± 2.82	68.56 ± 2.33	65.82 ± 3.13
		8-12-16-24	85.83 ± 2.81	65.66 ± 2.52	64.00 ± 3.10
		8-16-32-64	**89.96 ± 3.65**	67.56 ± 1.19	65.73 ± 0.73
	ResNet_{101}	4-8-16-32	74.79 ± 5.64	60.59 ± 9.34	54.34 ± 12.28
		6-12-24-48	87.18 ± 3.76	69.34 ± 1.26	66.74 ± 0.73
		8-12-16-24	84.40 ± 2.74	68.51 ± 2.89	65.95 ± 3.01
		8-16-32-64	83.28 ± 4.04	67.90 ± 1.89	64.77 ± 2.46
RetinaNet [11]	ResNet_{50}	4-8-16-32	73.06 ± 10.05	74.46 ± 4.75	67.66 ± 9.86
		6-12-24-48	74.05 ± 2.49	81.71 ± 1.52	76.47 ± 2.76
		8-12-16-24	78.86 ± 4.59	81.00 ± 1.23	75.09 ± 2.10
		8-16-32-64	80.26 ± 5.75	78.30 ± 3.56	71.37 ± 1.52
	ResNet_{101}	4-8-16-32	64.00 ± 6.44	67.77 ± 12.51	55.93 ± 14.06
		6-12-24-48	79.75 ± 3.75	78.74 ± 3.19	72.78 ± 3.32
		8-12-16-24	78.16 ± 1.89	80.14 ± 1.18	75.13 ± 1.21
		8-16-32-64	78.68 ± 2.16	**82.42 ± 0.52**	**76.83 ± 1.09**

TABLE V
DETECTION RESULTS OF VARYING 3D CNN ARCHITECTURES ON THE **HANDGUN** DATASET.

Model	Network	Anchor size	Score threshold=0.9		Average Precision (%)
			Precision (%)	Recall (%)	
Faster R-CNN [10]	ResNet_{50}	4-8-16-32	91.67 ± 1.82	84.62 ± 2.21	84.00 ± 2.07
		6-12-24-48	91.38 ± 4.19	86.98 ± 2.25	85.30 ± 3.11
		8-12-16-24	**92.43 ± 1.37**	86.38 ± 1.74	85.40 ± 1.77
		8-16-32-64	91.98 ± 3.33	87.56 ± 1.54	86.74 ± 1.81
	ResNet_{101}	4-8-16-32	89.93 ± 3.39	85.18 ± 4.74	83.98 ± 5.03
		6-12-24-48	91.00 ± 4.23	88.74 ± 1.76	87.76 ± 1.82
		8-12-16-24	91.70 ± 1.07	85.19 ± 2.31	84.38 ± 2.16
		8-16-32-64	90.17 ± 1.76	86.97 ± 3.41	85.92 ± 2.93
RetinaNet [11]	ResNet_{50}	4-8-16-32	87.29 ± 2.14	89.34 ± 1.52	87.30 ± 3.33
		6-12-24-48	90.06 ± 2.79	90.53 ± 0.88	**89.13 ± 0.87**
		8-12-16-24	91.15 ± 3.63	89.95 ± 1.58	88.12 ± 2.33
		8-16-32-64	88.98 ± 2.67	89.94 ± 0.91	85.89 ± 1.74
	ResNet_{101}	4-8-16-32	88.95 ± 2.21	90.53 ± 2.24	88.61 ± 2.25
		6-12-24-48	89.69 ± 4.30	**90.55 ± 2.13**	87.82 ± 2.65
		8-12-16-24	91.09 ± 1.33	90.54 ± 0.75	87.31 ± 1.35
		8-16-32-64	90.56 ± 2.10	90.54 ± 1.64	87.18 ± 2.40

icantly achieve better results (AP: 64.77 *Bottle* dataset, Fatser R-CNN with ResNet_{101}) compared to resampling factor of $1/2$ (AP: 58.68 *Bottle* dataset, Fatser R-CNN with ResNet_{101}). Therefore, the results reported in the Tables IV and V, the resampling factor of $1/3$ is applied with confidence score threshold of 0.9.

We vary the anchor sizes (4 different sets) in our experiments as explained in the Section III-B. The AP is higher with the larger anchor size, i.e., (*6-12-24-48*), (*8-16-32-64*) (AP: $\sim$65%), compared to anchor size of (*4-8-16-32*) (AP: $\sim$58%) while using Faster R-CNN [10] with ResNet_{50} (Table IV upper) for *Bottle* dataset. The similar trend is perceptible for both Faster R-CNN [10] and RetinaNet [11] (Table IV, upper and lower) with ResNet_{50} and ResNet_{101}. The best AP is achieved by RetinaNet with ResNet_{101} (AP: 76%, Table IV, lower) with anchor size of (*8-16-32-64*). For *Handgun* dataset, RetinaNet with ResNet_{50} achieves the highest average precision (AP: 89%, Table V, lower) using (*6-12-24-18*) anchor size. It observable that for *Handgun* dataset, all different variant of anchor sizes achieve similar performances on all the metrics (precision, recall and AP).

From the results (Tables IV, V), by increasing the number of convolutional layers in backbone network (ResNet_{50} vs ResNet_{101}) does not increase the performance. By comparing two different detection architectures, RetinaNet [11] outperforms Faster R-CNN [10] for both *Botte* dataset (AP: 76%, Table IV, lower) and *Handgun* dataset (AP: 89%, Table V, lower).

Exemplar prohibited items detection results from Faster R-CNN [10] and RetinaNet [11] with ResNet_{101} are depicted in Figure 4. From the examples, we observe Faster R-CNN [10] falsely detects a bottle while RetinaNet [11] correctly detects the item (Figure 4-Bottles, column 3). This is anticipated due to the superior performance of RetinaNet [11] than Faster R-CNN [10] for *Bottle* dataset (Table I). Both the models perform in a similar fashion for handguns as depicted in the Figure 4-Handguns, echoed our quantitative evaluations in Table II.

VI. CONCLUSION

We extend Convolutional Neural Networks for prohibited item classification and detection in volumetric 3D CT baggage

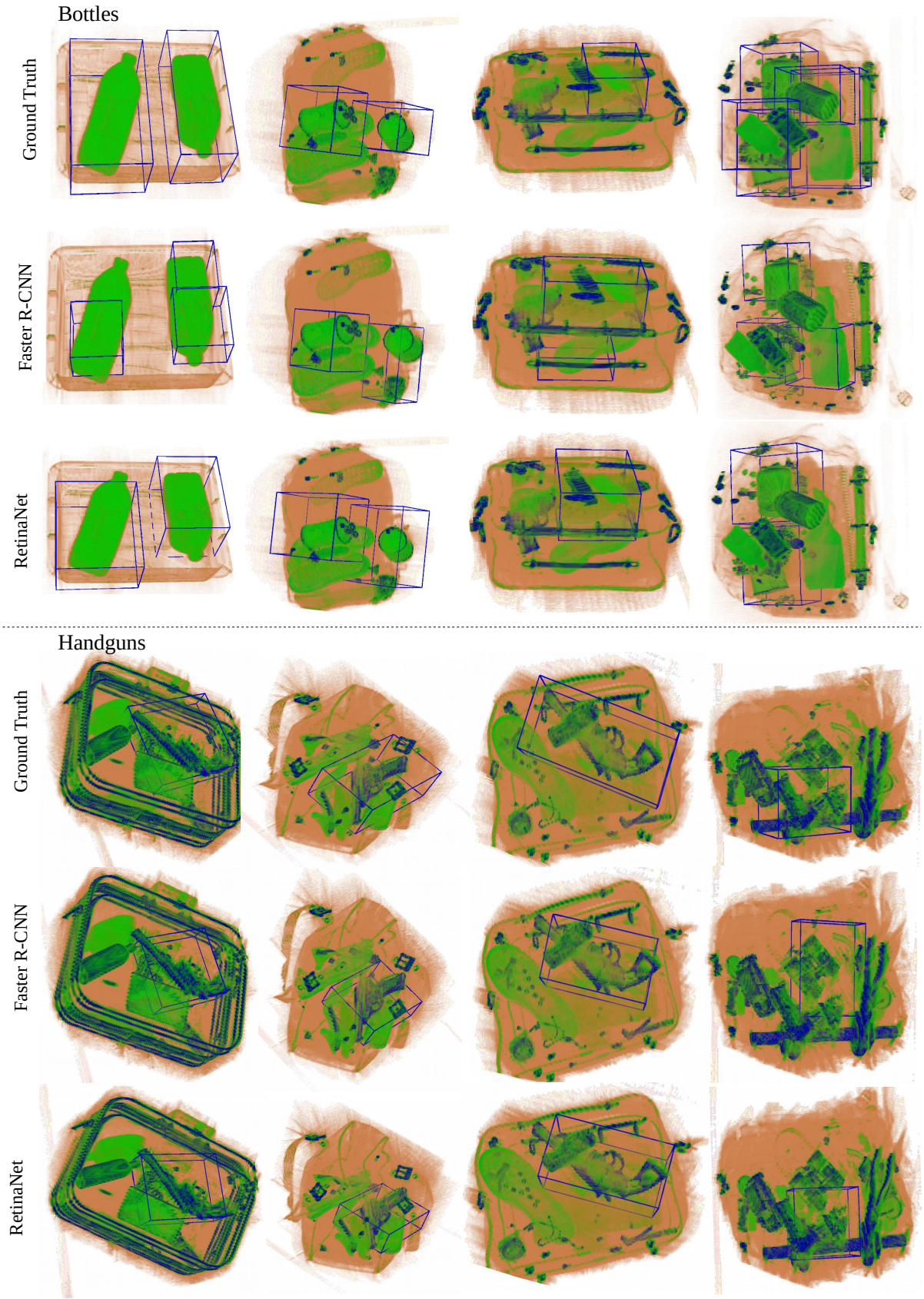

Fig. 4. Exemplar detection results (ground truth, detection results of Faster R-CNN [10] and RetinaNet [11] both with ResNet_{101} are shown in top three rows for bottles and bottom three rows for handguns).

security screening X-ray imagery. As the first attempt of deep CNN based techniques in this specific application, we make extensive evaluations on a variety of CNN models and data pre-processing strategies. The experimental results on classification demonstrate comparable performance (TP: $\sim$98%, FP: $\sim$1.5% for both *Bottle* and *Handgun*) of the Voxception-ResNet model with prior art using hand-crafted 3D features whilst the 3D CNN model is more computationally efficient than the traditional methods [14]. The results of detection demonstrate the feasibility of extending traditional 2D object detectors (e.g., Faster R-CNN and RetinaNet) to detect prohibited item in volumetric 3D CT data (mAP: $\sim$76% for *Bottle*, mAP: $\sim$88% for *Handgun*).

Although viable performance has been achieved for the classification task, the detection task still needs to be improved in order to meet operational requirements for aviation security screening. Limited training data is the primary reason for the more limited detection performance. In our future work, we will take advantage of transfer learning and the use of synthetic data [2] to improve the performance. In addition, we will incorporate variety of prohibited items for multi-class classification and detection problems.

REFERENCES

[1] N. Bhowmik, Q. Wang, Y. F. A. Gaus, M. Szarek, and T. P. Breckon, "The good, the bad and the ugly: Evaluating convolutional neural networks for prohibited item detection using real and synthetically composited X-ray imagery," in *British Machine Vision Conference Workshops*, 2019.

[2] Q. Wang, N. Megherbi, and T. P. Breckon, "A reference architecture for plausible threat image projection (TIP) within 3D X-ray computed tomography volumes," *Journal of X-ray Science and Technology*, 2020, in press.

[3] R. F. Meuter and P. F. Lacherez, "When and why threats go undetected: Impacts of event rate and shift length on threat detection accuracy during airport baggage screening," *Human factors*, vol. 58, no. 2, pp. 218–228, 2016.

[4] S. Akcay, 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. Gaus, N. Bhowmik, S. Akcay, and T. 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 Applications*. IEEE, December 2019.

[6] N. Bhowmik, Y. Gaus, S. Akcay, J. Barker, and T. Breckon, "On the impact of object and sub-component level segmentation strategies for supervised anomaly detection within x-ray security imagery," in *Proc. Int. Conf. on Machine Learning Applications*. IEEE, December 2019, to appear.

[7] N. Bhowmik, Y. Gaus, and T. Breckon, "Using deep neural networks to address the evolving challenges of concealed threat detection within complex electronic items," in *Proc. Conference on Homeland Security*. IEEE, November 2019, to appear.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. computer vision and pattern recognition*, 2016, pp. 770–778.

[9] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Generative and discriminative voxel modeling with convolutional neural networks," in *Neural Information Processing Systems*, 2016.

[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] T.-Y. 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. 2980–2988.

[12] D. F. Wiley, D. Ghosh, and C. Woodhouse, "Automatic segmentation of CT scans of checked baggage," in *Proc. Int. Meeting on Image Formation in X-ray CT*, 2012, pp. 310–313.

[13] G. Flitton, T. P. Breckon, and N. Megherbi, "A 3D extension to cortex like mechanisms for 3D object class recognition," in *Proc. Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3634–3641.

[14] A. Mouton, T. P. Breckon, G. T. Flitton, and N. Megherbi, "3D object classification in baggage computed tomography imagery using randomised clustering forests," in *Proc. Int. conf. on image processing (ICIP)*, 2014, pp. 5202–5206.

[15] P. Jin, D. H. Ye, and C. A. Bouman, "Joint metal artifact reduction and segmentation of CT images using dictionary-based image prior and continuous-relaxed potts model," in *Proc. Int. conf. on Image Processing (ICIP)*. IEEE, 2015, pp. 798–802.

[16] G. Flitton, A. Mouton, and T. P. Breckon, "Object classification in 3D baggage security computed tomography imagery using visual codebooks," *Pattern Recognition*, vol. 48, no. 8, pp. 2489–2499, 2015.

[17] A. Mouton and T. P. Breckon, "Materials-based 3D segmentation of unknown objects from dual-energy computed tomography imagery in baggage security screening," *Pattern Recognition*, vol. 48, no. 6, pp. 1961–1978, 2015.

[18] ——, "A review of automated image understanding within 3D baggage computed tomography security screening," *Journal of X-ray Science and Technology*, vol. 23, no. 5, pp. 531–555, 2015.

[19] Q. Wang, K. N. Ismail, and T. P. Breckon, "An approach for adaptive automatic threat recognition within 3D computed tomography images for baggage security screening," *Journal of X-ray Science and Technology*, vol. 28, no. 1, pp. 35–58, 2020.

[20] D. Maturana and S. Scherer, "Voxnet: A 3D convolutional neural network for real-time object recognition," in *Proc. Int. conf. on Intelligent Robots and Systems*. IEEE, 2015, pp. 922–928.

[21] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3D object detection," in *Proc. Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.

[22] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3D object detection from rgb-d data," in *Proc. Computer Vision and Pattern Recognition*, 2018, pp. 918–927.

[23] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang, "Automated pulmonary nodule detection in CT images using deep convolutional neural networks," *Pattern Recognition*, vol. 85, pp. 109–119, 2019.

[24] P. F. Jaeger, S. A. Kohl, S. Bickelhaupt, F. Isensee, T. A. Kuder, H.-P. Schlemmer, and K. H. Maier-Hein, "Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection," in *Proc. Neural Information Processing Systems Workshops*, 2019.

[25] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view cnns for object classification on 3D data," in *Proc. computer vision and pattern recognition*, 2016, pp. 5648–5656.

[26] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3D classification and segmentation," in *Proc. Computer Vision and Pattern Recognition*, 2017, pp. 652–660.

[27] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer learning for 3D medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019, unpublished.

[28] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. computer vision and pattern recognition*, 2014, pp. 580–587.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. computer vision and pattern recognition*, 2016, pp. 2818–2826.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. computer vision and pattern recognition*, 2016, pp. 779–788.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. conf. on Learning Representations*, 2015.