Contraband Materials Detection Within Volumetric 3D Computed Tomography Baggage Security Screening Imagery

Qian Wang
Department of Computer Science
Durham University
Durham, UK

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

***Abstract*—Automatic prohibited object detection within 2D/3D X-ray Computed Tomography (CT) has been studied in literature to enhance the aviation security screening at checkpoints. Deep Convolutional Neural Networks (CNN) have demonstrated superior performance in 2D X-ray imagery. However, there exists very limited proof of how deep neural networks perform in materials detection within volumetric 3D CT baggage screening imagery. We attempt to close this gap by applying Deep Neural Networks in 3D contraband substance detection based on their material signatures. Specifically, we formulate it as a 3D semantic segmentation problem to identify material types for all voxels based on which contraband materials can be detected. To this end, we firstly investigate 3D CNN based semantic segmentation algorithms such as 3D U-Net and its variants. In contrast to the original dense representation form of volumetric 3D CT data, we propose to convert the CT volumes into sparse point clouds which allows the use of point cloud processing approaches such as PointNet++ towards more efficient processing. Experimental results on a publicly available dataset (NEU ATR) demonstrate the effectiveness of both 3D U-Net and PointNet++ in materials detection in 3D CT imagery for baggage security screening.**

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

I. Introduction

Effective and efficient baggage screening at checkpoints in airports is crucial for aviation security. In most airports, X-ray machines are deployed to scan hand baggage for prohibited objects and contraband/threat materials. The reconstructed X-ray imagery is shown to human operators who will use their expertise and experience to find out potential prohibited items within the images. The task is challenging when the baggage is packed with clutter such as large electronics since the resulting inter-occlusion between objects may lead to difficulty in identifying potential threat and contraband items. For this reason, passengers are usually required to divest any large electronic devices (e.g., laptop, tablet) and liquids before security screening. 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.

In recent years, impressive progress in deep learning techniques has enabled the possibility of fully automatic prohibited object detection in 2D X-ray imagery with high precision and very low false alarm rates [1], [2]. With the success of automatic threat object detection in 2D X-ray imagery, attempts have been made to extend this idea to 3D CT imagery with promising results achieved in prior work [3], [4]. However, the techniques used in [3] and [4] rely on the detection of specific object appearance and shape (e.g., handguns, bottles, knives, etc.) hence is likely to fail in detecting contraband materials (e.g., explosive material, drugs, etc.) which can appear in arbitrary shapes. Existing research in contraband/threat material classification and detection are mainly based on traditional approaches such as morphological operations based segmentation followed by a classifier [5]. It is unknown how deep learning techniques perform in materials detection within 3D CT imagery.

To address this issue, we attempt to address the contraband material detection problem within volumetric 3D CT baggage security screening imagery. Specifically, we formulate it as a semantic segmentation problem and generate voxel-wise semantic labelling maps based on the materials. Post-processing is subsequently applied to the segmentation results to estimate the potential contraband material signatures. We use semantic segmentation methods such as the popular U-Net [6] architecture and its variants. Alternatively, we investigate the possibility of converting the dense volumetric 3D data to sparse point clouds and use point cloud processing methods (e.g., PointNet++ [7]) for semantic segmentation.

To evaluate the effectiveness of the proposed approach to contraband material detection in 3D CT volumes, we conduct experiments on the public Northeastern University Automatic Threat Recognition (NEU ATR) dataset [8]. Experimental results demonstrate our proposed approach using 3D Convolutional Neural Networks (CNN) based U-Net architectures outperforms traditional 3D segmentation based methods [5] whilst point cloud based methods are more computationally efficient.

To summarize, the contributions of this paper are as follows:

- the first attempt to address contraband materials detection within volumetric 3D CT baggage security screening imagery using various deep learning models such as 3D U-Net and PointNet++.
- the first attempt to convert volumetric 3D data to point clouds to promote computationally efficient processing and to compare 3D U-Net architectures and PointNet++ as two differing representation paradigms for volumetric 3D imagery processing.
- a framework for contraband materials detection is proposed by formulating it as a semantic segmentation problem followed by post-processing operations and is validated through extensive experiments on a publicly available dataset with three types of target materials for detection and classification (i.e. saline, rubber and clay).

II. Related Work

In this section, we review existing work related to our own from the perspective of *baggage security screening* and *3D semantic segmentation*.

A. Baggage Security Screening

Automatic object detection and recognition algorithms have been proposed and evaluated for baggage aviation security screening based on 2D X-ray images [1], [9]. The use of CNN architectures and object detection frameworks boosts the performance with a high detection rate and a low false-positive rate. For instance, Gaus et al. [10] evaluate the effectiveness of Faster R-CNN [11], Mask R-CNN [12] and RetinaNet [13] in detecting six different objects (i.e. bottle, hairdryer, iron, toaster mobile and laptop) in 2D X-ray baggage images.

To enable automatic baggage screening using 3D CT imagery, a variety of studies have been carried out in recent years [5], [14]–[21].

One research direction is object segmentation based on the material and morphological structure [5], [14], [19]. Specifically, Mouton et al. [19] propose 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. [5] along with others [8], [22] 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 is proposed based on recursive morphological operations and the Support Vector Machines (SVM) were employed for the classification of three types of materials [5].

3D object detection within 3D CT baggage security screening imagery has been studied in [3], [23], [24]. Flitton et al. [24] evaluate the effectiveness of different 3D descriptors in a search-based detection approach. Their approach is limited to detect known objects for which the reference data are assumed to be available. Such an assumption hinders its application in practice when the reference data are usually unavailable. Wang et al. [3] use contemporary object detection frameworks based on 3D CNN and evaluate its performance on individual object detection independently. However, most of these existing works focus on the prohibited objects of specific appearances and shapes and may not perform well for contraband items of specific materials. To close this gap, we attempt to investigate the possibility of using 3D deep learning techniques [25], [26] for contraband materials detection within volumetric 3D CT imagery for security screening. In parallel to the research mentioned above, there also exist studies on Explosive Device Systems (EDS) for aviation security screening [27]–[30] which focus on the detection of explosive materials.

B. 3D Segmentation

3D segmentation is a typical approach to 3D scene understanding based on RGB-D data [31]–[34]. The first category of approaches to RGB-D semantic segmentation encoded the depth map as an image which can be processed by 2D CNN in a similar way to RGB image processing [31], [33]. Alternatively, 3D CNN models were employed in [32], [34]. The depth maps were used to represent the 2D RGB images into the corresponding 3D volumetric representations. By comparison, the CT data are naturally in the form of 3D volumes. However, 3D CNN models suffer from dealing with high-resolution data due to the computation cost. To alleviate this issue, 3D graph neural networks were proposed for RGBD semantic segmentation [35].

Medical image analysis is one of the most active areas of 3D segmentation. U-Net [6] is an effective architecture of deep neural network model for semantic segmentation. It has been extended to its 3D variant for a variety of applications such as pulmonary nodule segmentation [36], [37], skin lesion segmentation [38], kidney tumor segmentation [39] and infant brain segmentation [40] in medical images of varying modalities (e.g., Electron Microscopic, CT and MRI). Specifically, MultiResUNet [38] modifies the original U-Net by employing Inception-style blocks to capture multi-scale features in the contraction path (i.e. the encoder). A similar idea of Inception modules was also employed by [40] for infant brain MRI segmentation. Residual U-Net introduces residual modules (i.e. skip connections [41]) to the CNN blocks in the U-Net architecture and was reported to outperform U-Net by [39]. A transfer learning framework was proposed in [37] to alleviate the training data sparsity issues of most medical image analysis tasks.

Other than the applications in medical image analysis, 3D semantic segmentation has also been extensively studied in the domain of point cloud data analysis (e.g., Lidar point cloud) for a variety of applications including autonomous driving. PointNet [42] and its variant PointNet++ [7] are two leading contemporary methods using end-to-end frameworks directly encoding the point cloud into context-aware features for dif-

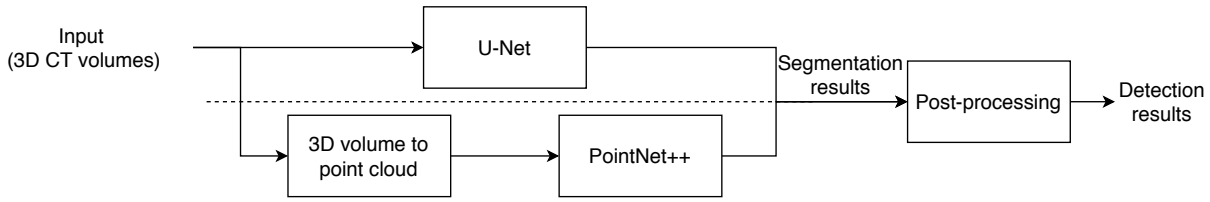

Fig. 1. The pipeline of our approach to contraband materials detection in baggage CT volumes (two options for semantic segmentation are represented as two branches, see details in Section III-A and III-B).

ferent down-stream tasks including 3D semantic segmentation. Due to their generality and effectiveness demonstrated in the literature, we employ both PointNet and PointNet++ to investigate the possibility of converting 3D CT volumes to point clouds for efficient semantic segmentation. Other alternative approaches to point cloud segmentation use the idea of voxelization to transform the sparse point clouds to grid representations which can be fed into 3D CNN models [43]. This is an inverse process of our method and has the limitation of high computation cost. Since the raw data we are concerned with is in the form of 3D volumes, the other alternative to reducing the computation cost is to use sparse convolutional networks [44] which is also employed by [43]. In our work, we focus mainly on the performance of different semantic segmentation methods and hence we use a simple down-sampling strategy to reduce image resolution for more efficient computation.

III. Method

Our work focuses on the detection of contraband materials with no specific appearances/shapes. Existing prohibited object detection works focus on the detection of prohibited objects such as firearms and knives employ traditional object detection frameworks (i.e. predicting a bounding box for each detected object) [3]. However, this is not an optimal choice for the detection of contraband materials which can be of arbitrary shapes (e.g., curved sheets, liquid in different containers). As a result, we formulate it as a *semantic segmentation* problem and attempt to predict voxel-wise labels for 3D baggage CT volumes. The segmentation results give potential locations of contraband items and the classes they belong to. Post-processing is subsequently employed to refine the segmentation results and generate contraband materials detection results.

In the following subsections, we first introduce two approaches to 3D semantic segmentation. As we formulate it as a semantic segmentation problem, we first extend the prevalent U-Net architecture [6], [45] for 2D image segmentation to our 3D CT segmentation scenario. To improve efficiency and reduce the memory and processing time consumption, we also explore the possibility of using point cloud processing methods [7], [42] for CT segmentation. To this end, we convert the volumetric 3D CT data into sparse point clouds which reserve only a small fraction of useful voxels in the original volumes as points. As illustrated in Figure 1, segmentation results are finally post-processed to generate contraband materials detection results.

A. 3D CNN Based Method

We follow the work in [45] and extend U-Net architecture to 3D scenarios. As shown in Figure 2, 3D U-Net also consists of a contraction path and an expanding path. The contraction path is composed of a sequence of down-sampling modules to capture context from the input 3D volumes whilst the expanding path uses a sequence of upsampling modules to expand the low-resolution feature volumes extracted by the contraction path to the original resolution. There are L down-sampling modules in the contraction path and the same number of up-sampling modules in the expanding path. Skip connections are used to copy the feature volumes from the contraction path and concatenate them with the feature volumes of the same level in the expanding path.

The down-sampling module in 3D U-Net architecture is composed of two 3D convolution layers, each of which is followed by ReLU and batch normalisation layers. A pooling layer is used to down-sampling the spatial resolution of feature volumes by a factor of 2 in all three dimensions. Specifically, the number of feature volumes in the l-th level is 2^l times more than that in the 0-th level (denoted as $fvols$) whilst the dimension of the feature volumes is $1/2^l$ of that in the original CT volume. On the other hand, the up-sampling module is implemented by a non-parametric upsample function in 3D U-Net followed by two consecutive 3D convolution layers (each is followed by ReLU and batch normalisation layers) similar to the counterparts in the contraction path. The details of down-sampling and up-sampling modules are shown in the dashed boxes of Figure 2.

Figure 2 also illustrates the details of down-sampling and up-sampling modules of the residual 3D U-Net architecture in the dashed boxes. Compared with those in 3D U-Net, the main difference is three-fold. Firstly, there are three 3D convolution layers (as well as the ReLU and batch normalisation layers) in each module. Secondly, there is a skip connection between the first and third convolution layers. Finally, a 3D deconvolution layer is used in residual 3D U-Net as opposed to the interpolation-based up-sampling layer in the 3D U-Net.

B. Point Cloud Based Methods

In a baggage CT volume, there are usually large regions of non-threat voxels which can be easily recognized by simple thresholding based on prior knowledge of the density ranges

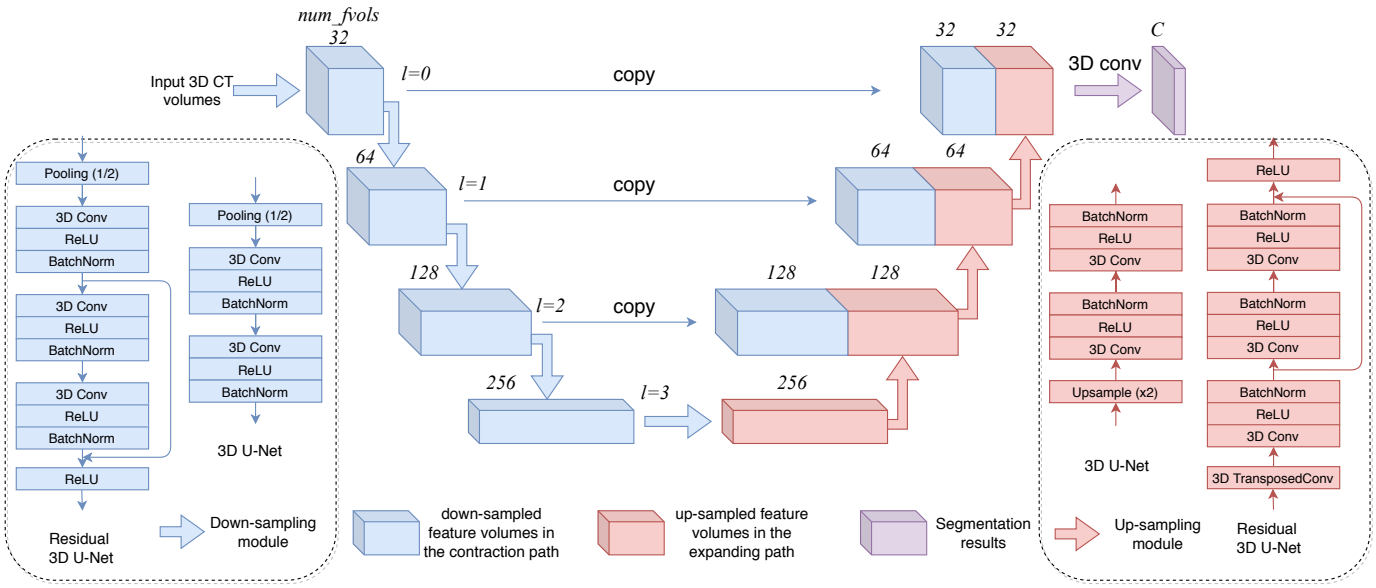

Fig. 2. 3D U-Net architectures for semantic segmentation.

of these benign materials in Hounsfield units. To alleviate the memory and time-intensive issues of 3D CNN models, we attempt to investigate the possibility of point cloud processing algorithms in semantic segmentation for volumetric 3D CT data.

In the first stage, CT volumetic data is converted into a point cloud by reserving only the voxel-of-interest. Specifically, we use prior knowledge to set thresholds and consider only voxels whose intensity values are within a specified range where contraband materials fall into. As a result, a point is represented by a 4-dimensional vector of coordinates x, y, z and the intensity i.

We use the most popular point cloud processing models PointNet [42] and its extension PointNet++ [7] for the proof of concept as they perform the best among a few candidates in our preliminary experiments. PointNet and PointNet++ take point clouds as the input and generate segmentation results for our purpose. We use the default model architecture proposed in the original paper.

Specifically, PointNet takes n points from a point cloud (may be a sub-block within a large point cloud) as input which is represented as a $n \times 4$ matrix. PointNet first transforms the input by a learnable 3×3 transformation matrix. Subsequently, all point features are fed into a Multi-Layer Perceptron module (MLP) and transformed to a $n \times 64$ feature matrix. Similar processing is repeated and finally a $n \times 1024$ feature matrix is generated and max-pooled to get a global feature of 1024 dimensions. For semantic segmentation in our case, the global feature is concatenated with each point feature (i.e. the one of 64 dimensions) to form a n feature matrix which again is fed into a sequence of MLP until the final output layer generating the segmentation results.

PointNet++, as an extension of PointNet, takes n points from a point cloud as input which is represented as a $n \times 7$ matrix where the three more features than those used in PointNet are normalized point coordinates x', y' and z'. PointNet++ learns hierarchical point set features via the set abstraction module which is composed of point sampling and grouping followed by PointNet based feature extraction. The set abstraction modules are repeated for twice before a sequence of interpolation and unit PointNet to generate final segmentation results of the same resolution as the input.

C. Post-processing

The segmentation results of 3D U-Net and PointNet++ are voxel-wise and point-wise class labelling respectively. We convert these segmentation results to detection results in the post-processing stage. Specifically, we group the connected voxels which are labelled as the same class as a detected object. To these ends, we use morphological operations to correct the mislabelling information in the segmentation results.

The pipeline of post-processing is shown in Figure 3. For each foreground class, we apply *dilation* and *erosion* operations sequentially to the binary segmentation map to correct the missing voxel labels within the detected objects. Subsequently, the connected component labelling (CCL) algorithm is employed to group the labelled voxels into a set of potential detected objects. We prune the detection results by removing the objects whose volumes are smaller than a pre-defined threshold.

IV. Experiments and Results

In this section, we conduct experiments on a public baggage CT dataset to evaluate the effectiveness of our proposed approaches to contraband materials detection. We describe the details of the dataset and evaluation metrics used in our experiments. Subsequently, we report the quantitative results of 3D CNN methods and point cloud methods respectively.

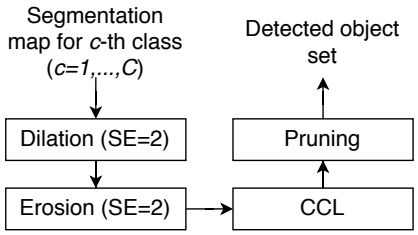

Fig. 3. The pipeline of post-processing using morphological operations to convert segmentation results to detection results.

Finally, qualitative evaluation is presented to provide some intuitive insight into how our approaches perform.

A. Dataset

We follow [5] and use the Northestern University Automated Threat Recognition (NEU ATR) dataset [8], [46] collected and annotated by NEU ALERT throughout our experiments in this study. Baggage CT volumes were collected by a medical CT scanner (Imatron C-300). The slice size is 512×512 corresponding to the field view of 475 mm×475 mm hence the in-plane pixel size is 0.928 mm. The number of slices varies in different volumes and the slice spacing is 1.5 mm. Pixel values are represented by the Modified Hounsfield Unit (MHU) ranging from 0 to 32,767 MHU in which air and water are 0 and 1024 respectively.

The ATR dataset consists of 188 CT volumes in which there are 446 object signatures of three target materials (i.e. *saline*, *rubber* and *clay*) and other non-target materials as cluttered background of typical packed baggage. The ground truth voxels are labelled by NEU ALERT for all the objects of three target materials. We follow [5] to split the whole dataset into two subsets evenly: *odd* set and *even* set containing 94 odd and even indexed volumes respectively (i.e. 50/50, training/testing data split). In our experiments, we use one subset for training and the other for testing.

B. Evaluation Metrics

We use three groups of evaluation metrics in our experiments. The first one is the mean Intersection over Union (IoU) which is a typical evaluation metric for semantic segmentation as the contraband materials detection has been formulated as a semantic segmentation problem. IoU is computed for each class and the mean IoU is the mean of each IoU for all classes. The mean IoU evaluates the performance of segmentation but cannot measure the detection performance of individual objects in a CT volume.

The second group of evaluation metrics are the precision and recall which can be computed based on the detection results obtained after post-processing the segmentation results (c.f. Section III-C).

To make a direct comparison with the traditional method proposed in [5], we also use a third group of evaluation metrics in our experiments which have been used in [5]. These evaluation metrics are similar to typical ones for object detection (i.e. precision and recall in the second group) but concern the Probability of Detection (PD) and the Probability of False Alarms (PFA). PD is similar to recall and the main difference is that PD is computed over all detections regardless of their classes whilst recall is computed class-wisely. PD is defined in this way so that the detection model focuses more on the difference between contraband items and benign ones rather than the difference between different types of contraband items.

TABLE I
IOU RESULTS OF MATERIAL SEGMENTATION WITHIN 3D CT VOLUMES ON NEU ATR DATASET (L – # OF LEVELS IN 3D U-NET; FVOLS – # OF FEATURE VOLUMES IN 3D U-NET; FAC. – THE DOWNSAMPING FACTOR USED TO DOWN-SAMPLE THE ORIGINAL CT VOLUMES).

Method	L	fvols	fac.	Background	Saline	Rubber	Clay	Overall
PointNet	-	-	2	99.0	15.7	15.5	31.0	40.3
PointNet	-	-	4	99.1	3.9	5.1	35.5	35.9
PointNet++	-	-	2	98.9	39.8	28.2	61.9	57.2
PointNet++	-	-	4	99.0	32.6	26.9	50.9	52.3
3D U-Net	4	32	2	99.5	65.3	58.8	74.5	74.5
3D U-Net	4	32	4	99.5	64.2	61.3	66.5	72.9
3D U-Net	4	32	8	99.5	58.7	48.7	60.7	66.9
3D U-Net	4	64	8	**99.6**	61.8	53.3	64.5	69.8
3D U-Net	6	32	4	**99.6**	65.5	62.1	69.2	74.1
3D U-Net	6	64	4	**99.6**	64.9	63.0	72.5	**75.0**
Resisual 3D U-Net	4	32	2	99.5	56.6	57.7	**74.8**	72.2
Resisual 3D U-Net	4	32	4	**99.6**	63.1	60.9	67.0	72.6
Resisual 3D U-Net	6	32	4	**99.6**	63.5	59.4	73.1	73.9
Resisual 3D U-Net	6	64	4	**99.6**	**67.4**	**64.6**	67.9	74.9

C. Experimental Settings

Extensive experiments are conducted to evaluate the effectiveness of the proposed methods and to investigate how different factors affect the performance.

1) Model architecture: We investigate two types of 3D CNN models (i.e. 3D U-Net and residual 3D U-Net) and one point cloud model (i.e. PointNet++) in our experiments. For 3D U-Net and residual 3D U-Net architectures, we consider varying depth and width by setting the number of feature volumes in the first level (l=0) as $fvols \in \{32, 64\}$ and the number of down-sampling and up-sampling modules as $L \in \{4, 6\}$ based on the Pytorch implementation of [45]. As a result, we can have a number of candidate models with varying combinations of settings. Instead of investigating all the possible combinations, we select typical ones performing better as shown in Tables I-III and ignore those providing less insight. In all experiments, we set the learning rate to $1e-4$ and decrease it every 50 epochs by a factor of 0.5. The training is terminated after 250 epochs. During training, we randomly crop 3D sub-volumes of size $64 \times 96 \times 96$ and apply data augmentation including normalisation, random flipping and random rotation (by 90 degrees). Each training batch constitutes of 16 such sub-volumes from cropped 4 baggage volumes (4 from each).

For PointNet and PointNet++, we use the default settings in the PyTorch implementation [47] except the input feature dimension is adapted since we have only one intensity feature in our CT data as opposed to three RGB values. The learning rate is set to $1e-3$ throughout our experiments and decays by a factor of 0.7 every 50 epochs. The training stops after 250 epochs. In the training, we randomly select 8092 points

TABLE II

PRECISION AND RECALL RESULTS OF MATERIAL SEGMENTATION WITHIN 3D CT VOLUMES ON NEU ATR DATASET (L – # OF LEVELS IN 3D U-NET; FVOLS – # OF FEATURE VOLUMES IN 3D U-NET; FAC. – THE DOWNSAMPING FACTOR USED TO DOWN-SAMPLE THE ORIGINAL CT VOLUMES).

Model	L	fvols	fac.	Saline		Rubber		Clay		Overall		
				P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	F1 (%)
PointNet	-	-	2	35.0	62.8	40.4	56.8	58.2	73.6	44.5	64.4	52.6
PointNet	-	-	4	31.1	16.2	27.6	13.6	54.5	70.6	37.8	33.5	35.5
PointNet++	-	-	2	41.9	84.1	59.7	58.9	60.1	76.9	53.9	73.3	62.1
PointNet++	-	-	4	37.8	73.2	56.8	51.3	52.5	79.8	49.0	68.1	57.0
3D U-Net	4	32	2	59.1	**94.6**	82.6	85.8	76.5	**94.2**	72.7	**91.5**	81.0
3D U-Net	4	32	4	77.1	85.2	81.3	90.3	83.7	86.6	80.7	87.3	83.9
3D U-Net	4	32	8	72.4	80.3	82.2	80.6	83.8	80.2	79.5	80.4	79.9
3D U-Net	6	32	4	74.6	90.1	86.2	91.8	86.8	90.5	82.6	90.8	86.5
3D U-Net	6	64	4	71.8	91.0	87.4	**92.5**	**94.6**	85.8	84.6	89.8	**87.1**
3D Residual U-Net	4	32	2	67.2	79.1	82.2	80.6	79.4	92.6	76.3	84.1	80.0
3D Residual U-Net	4	32	4	78.0	82.8	83.8	88.1	93.2	83.0	85.0	84.7	84.8
3D Residual U-Net	6	32	4	**79.5**	79.3	84.8	89.1	89.2	88.0	84.5	85.5	85.0
3D Residual U-Net	6	64	4	78.4	82.5	**87.8**	87.3	90.4	91.3	**85.5**	87.1	86.3

TABLE III

PD AND PFA RESULTS OF MATERIAL SEGMENTATION WITHIN 3D CT VOLUMES ON NEU ATR DATASET (L – # OF LEVELS IN 3D U-NET; FVOLS – # OF FEATURE VOLUMES IN 3D U-NET; FAC. – THE DOWNSAMPING FACTOR USED TO DOWN-SAMPLE THE ORIGINAL CT VOLUMES).

Model	L	fvols	fac.	PD (%)				PFA (%)
				Saline	Rubber	Clay	Overall	Overall
SVM [5]	-	-	-	87	95	**96**	92	24
PointNet	-	-	2	81	84	88	84	29
PointNet	-	-	4	38	41	80	50	13
PointNet++	-	-	2	**97**	87	94	92	24
PointNet++	-	-	4	92	80	86	86	22
3D U-Net	4	32	2	95	94	92	**94**	11
3D U-Net	4	32	4	86	96	86	90	6
3D U-Net	4	32	8	81	84	81	82	5
3D U-Net	6	32	4	91	96	89	92	5
3D U-Net	6	64	4	91	**97**	83	91	6
3D Residual U-Net	4	32	2	75	85	86	82	7
3D Residual U-Net	4	32	4	85	94	83	88	5
3D Residual U-Net	6	32	4	80	94	92	89	5
3D Residual U-Net	6	64	4	86	92	92	90	**4**

from a set of points within a block of size $48 \times 48 \times 48$. The key to training PointNet and PointNet++ is to balance the training samples of different classes. Since the background points are the majority in the training data, we need to give more weights to points belonging to foreground classes during selection (i.e. make it more likely to select the block containing foreground class points). To this end, we tend to select the blocks containing more than 50% foreground class points as training samples.

2) Data down-sampling: We also investigate how the down-sampling factor affects the performance of our methods. Specifically, we down-sample the CT volumes uniformly by the factor of 2, 4 and 8 in all three dimensions respectively. During training, the ground truth labelling volumes are correspondingly down-sampled. In the evaluation, to make different results comparable, we up-sample the predicted low-resolution results to the original size and calculate the evaluation metrics.

D. Experimental Results

As the key component of our proposed approach to contraband materials detection, different semantic segmentation models are evaluated by calculating the per-class IoU and mean IoU. As intermediate results, the performance of segmentation determines the final detection performance to a large extent. The IoU results of different models are reported in Table I. The majority of voxels belong to the background class hence the IoU is close to 100% for all methods. For three foreground classes, 3D U-Net architectures outperform point cloud based methods significantly with the best mean IoU of 75.0%. The use of skip connection in residual 3D U-Net models does not make a difference from those without skip connections. Increasing the width and depth of the architectures of 3D U-Net benefit the performance consistently. For point cloud based methods, PointNet++ outperforms PointNet significantly. The best point cloud based method is PointNet++ with the down-sampling factor of 2 and achieves the mean IoU of 57.2%.

Detection results are evaluated after post-processing by calculating the metrics of precision/recall. The results are shown in Table II and Table III respectively. Similar conclusions can be drawn from results in Table II and Table I. The best overall precision and recall is achieved by 3D U-Net with the depth of 6 and 64 feature volumes in the first level and the overall precision and recall are 84.6% and 89.8% respectively.

Following previous works in [5], we report the results of PD and PFA in Table III. By comparing with the results of [5], both the point cloud based methods and 3D U-Net based methods can achieve comparable or better performance. Although PointNet gives worse results than the traditional method, its variant PointNet++ can achieve comparable PD of 92% and PFA of 24%. Consistent with previous results, 3D U-Net with the depth of 6 and 32 feature volumes in the first level gives the best performance with the PD of 92% and a much less PFA of 5%.

E. Qualitative Evaluation

Figure 4 presents five exemplar samples of CT volumes from the ATR dataset together with their ground truth labelling and segmentation results generated by the best PointNet++ and 3D U-Net models in our experiments. The first two columns (from the left side) list the visualisations of CT volumes and their corresponding ground truth labelling. Three types of materials saline, rubber and clay are represented by *orange*, *green* and *blue* colours respectively. The last two columns list the segmentation results of PointNet++ and 3D U-Net

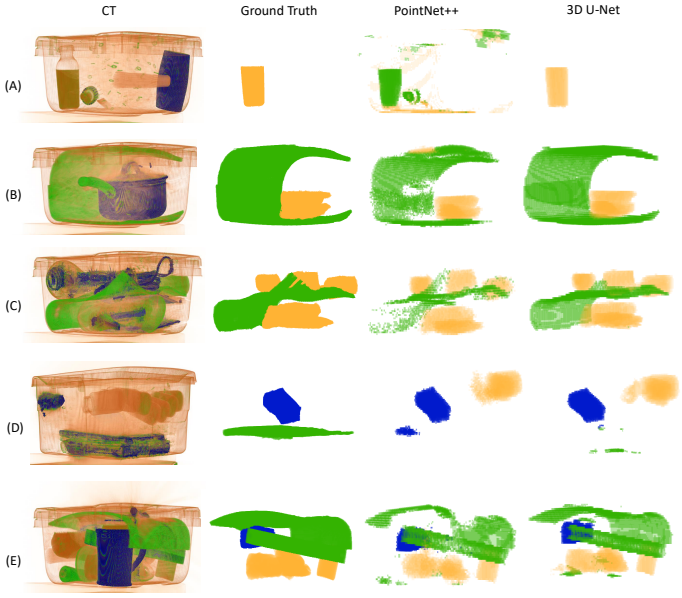

Fig. 4. Qualitative evaluation of material segmentation and classification using varying methods for examples A-E.

respectively. It can be seen from example (A) that some background voxels are mistakenly classified as foreground classes which will lead to low precision in Table II and high PFA in Table III. In addition, PointNet++ mistakenly classifies the material *saline* as *rubber*. The examples (B) and (C) shows that PointNet++ misses a considerable amount of voxels belonging to a *rubber* sheet (green) but 3D U-Net gives much better results. There also exist cases where both PointNet++ and 3D U-Net fail to detect the rubber sheet in example (D) but mistakenly detect a false alarm *saline* object.

F. Computational Complexity

Table IV presents the computational complexity of different models investigated in our work. We consider the number of parameters and floating point operations (FLOP) given a typical 3D CT volume with a size of $300 \times 512 \times 512$ in NEU ATR dataset. From Table IV, we can draw the following conclusions. Firstly, PointNet++ not only performs better but also has fewer parameters and computational cost than PointNet. Secondly, 3D U-Net is more efficient than 3D residual U-Net since it has fewer parameters but comparable performance when the depth and width are the same. Finally, increasing the depth and width of 3D U-Net is not efficient given the marginal performance gain and significantly increased computational cost. All these conclusions provide insightful instructions for our future work on volumetric 3D CT segmentation.

V. Conclusion

In this work, we have investigated two deep learning methods for contraband materials detection in volumetric 3D CT imagery. It is demonstrated that both 3D CNN and point cloud based methods can give reasonably good results for this task and 3D U-Net outperforms PointNet++ in terms of varying evaluation metrics. However, the point cloud based methods provide an alternative solution to the efficient processing of large 3D CT volumes and are worthy of further investigations.

In the future, we would like to expand the evaluation dataset and consider more types of contraband materials in aviation security screening. Besides, the detection performance can be enhanced by employing more advanced architectures [48] based on either 3D U-Net or point cloud processing algorithms. Finally, it will be of great value to integrate the detection of the prohibited object based on appearances [3] and contraband items based on materials in a unified framework for plausible real-world applications.

TABLE IV
Computational complexity of different models.

Model	L	fvols	fac.	# Parameters (M)	FLOP (G)
PointNet	-	-	2	3.53	3440
PointNet	-	-	4	3.53	430
PointNet++	-	-	2	**0.97**	460
PointNet++	-	-	4	**0.97**	57
3D U-Net	4	32	2	4.08	2250
3D U-Net	4	32	4	4.08	280
3D U-Net	4	32	8	4.08	**35**
3D U-Net	6	32	4	51.86	305
3D U-Net	6	64	4	207.43	1220
3D Residual U-Net	4	32	2	8.77	3520
3D Residual U-Net	4	32	4	8.77	440
3D Residual U-Net	6	32	4	84.87	470
3D Residual U-Net	6	64	4	339.44	1880

Acknowledgement [8], [46] - This material is based upon work supported by the U.S. Department of Homeland Security, Science and Technology Directorate, Office of University Programs, under Grant Award 18STEXP00001-03-02, Formerly 2013-ST-061-ED0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

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] 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.

[3] Q. Wang, N. Bhowmik, and T. P. Breckon, "On the evaluation of prohibited item classification and detection in volumetric 3d computed tomography baggage security screening imagery," in *International Joint Conference on Neural Networks*, 2020, to appear.

[4] ——, "Multi-class 3D object detection within volumetric 3D computed tomography baggage security screening imagery," in *Proc. IEEE Int. Conf. on Machine Learning and Applications*, 2020.

[5] 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.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[7] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*, 2017, pp. 5099–5108.

[8] C. Crawford, "Advances in automatic threat recognition (ATR) for CT-based object detection systems," https://myfiles.neu.edu/groups/ALERT/strategic_studies/TO4_FinalReport.pdf, 2015.

[9] 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.

[10] Y. F. A. Gaus, N. Bhowmik, S. Akçay, P. M. Guillén-Garcia, J. W. Barker, and T. P. Breckon, "Evaluation of a dual convolutional neural network architecture for object-wise anomaly detection in cluttered x-ray security imagery," in *International Joint Conference on Neural Networks*. IEEE, 2019, pp. 1–8.

[11] 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.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision*. IEEE, Oct 2017, pp. 2980–2988.

[13] 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.

[14] 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.

[15] 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.

[16] 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. International Conference on Image Processing*, 2014, pp. 5202–5206.

[17] 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 *Proceedings of International Conference on Image Processing*. IEEE, 2015, pp. 798–802.

[18] 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.

[19] 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.

[20] ——, "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.

[21] 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*, vol. 28, no. 3, pp. 507–526, 2020.

[22] D. W. Paglieroni, H. Chandrasekaran, C. Pechard, and H. E. Martz, "Consensus relaxation on materials of interest for adaptive ATR in CT images of baggage," in *Anomaly Detection and Imaging with X-Rays III*, vol. 10632. International Society for Optics and Photonics, 2018, p. 106320E.

[23] N. Megherbi, G. T. Flitton, and T. P. Breckon, "A classifier based approach for the detection of potential threats in ct based baggage screening," in *International Conference on Image Processing*. IEEE, 2010, pp. 1833–1836.

[24] G. Flitton, T. P. Breckon, and N. Megherbi, "A comparison of 3d interest point descriptors with application to airport baggage object detection in complex ct imagery," *Pattern Recognition*, vol. 46, no. 9, pp. 2420–2436, 2013.

[25] 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.

[26] 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.

[27] P. M. Edic, M. E. Vermilyea, F. F. Hopkins, G. Harding, and P. Landolfi, "Integrated multi-sensor systems for and methods of explosives detection," Jan. 11 2011, uS Patent 7,869,566.

[28] K. Wells and D. Bradley, "A review of X-ray explosives detection techniques for checked baggage," *Applied Radiation and Isotopes*, vol. 70, no. 8, pp. 1729–1746, 2012.

[29] C. C. Chan, A. Ferworn, and L. Chin, "Towards determining relative densities for common unknown explosives in improvised explosive devices," in *IEEE Canada International Humanitarian Technology Conference*. IEEE, 2017, pp. 55–60.

[30] D. Jumanazarov, J. Koo, M. Busi, H. F. Poulsen, U. L. Olsen, and M. Iovea, "System-independent material classification through X-ray attenuation decomposition from spectral X-ray ct," *NDT & E International*, p. 102336, 2020.

[31] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conf. on Computer Vision*. Springer, 2014, pp. 345–360.

[32] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 1912–1920.

[33] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, "LSTM-CF: Unifying context modeling and fusion with lstms for RGB-D scene labeling," in *European Conf. on Computer Vision*. Springer, 2016, pp. 541–557.

[34] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 1746–1754.

[35] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3D graph neural networks for RGB-D semantic segmentation," in *Proc. IEEE Int. Conf. on Computer Vision*, 2017, pp. 5199–5208.

[36] W. Funke, B. Veasey, J. Zurada, H. Frigui, and A. Amini, "3D U-Net for segmentation of pulmonary nodules in volumetric CT scans from multi-annotator truth estimation," in *Medical Imaging 2020: Computer-Aided Diagnosis*, vol. 11314. Int. Society for Optics and Photonics, 2020, p. 1131429.

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

[38] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[39] F. Isensee and K. H. Maier-Hein, "An attempt at beating the 3D U-Net," *arXiv preprint arXiv:1908.02182*, 2019.

[40] S. Qamar, H. Jin, R. Zheng, P. Ahmad, and M. Usama, "A variant form of 3D-UNet for infant brain segmentation," *Future Generation Computer Systems*, vol. 108, pp. 613–623, 2020.

[41] 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.

[42] 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.

[43] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "RandLA-Net: Efficient semantic segmentation of large-scale point clouds," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 11 108–11 117.

[44] B. Graham, M. Engelcke, and L. Van Der Maaten, "3D semantic segmentation with submanifold sparse convolutional networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 9224–9232.

[45] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432.

[46] ALERT, "Automated threat recognition (ATR) initiative dataset," http://www.northeastern.edu/alert/transitioning-technology/automated-threat-recognition-atr-initiative/, 2015.

[47] X. Yan, "Pytorch implementation of PointNet and PointNet++," https://github.com/yanx27/Pointnet_Pointnet2_pytorch, 2019.

[48] X. Yan, C. Zheng, Z. Li, S. Wang, and S. Cui, "PointASNL: Robust point clouds processing using nonlocal neural networks with adaptive sampling," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 5589–5598.