A deep learning framework for the automated inspection of complex dual-energy x-ray cargo imagery

Thomas W. Rogers[a,c], Nicolas Jaccard[a], and Lewis D. Griffin[a]

[a]Department of Computer Science, University College London, UK
[c]Department of Security and Crime Science, University College London, UK

ABSTRACT

Previously, we investigated the use of Convolutional Neural Networks (CNNs) to detect so-called Small Metallic Threats (SMTs) hidden amongst legitimate goods inside a cargo container. We trained a CNN from scratch on data produced by a Threat Image Projection (TIP) framework that generates images with realistic variation to robustify performance. The system achieved 90% detection of containers that contained a single SMT, while raising 6% false positives on benign containers. The best CNN architecture used the raw high energy image (single-energy) and its logarithm as input channels. Use of the logarithm improved performance, thus echoing studies on human operator performance. However, it is an unexpected result with CNNs.

In this work, we (i) investigate methods to exploit material information captured in dual-energy images, and (ii) introduce a new CNN training scheme that generates 'spot-the-difference' benign and threat pairs on-the-fly. To the best of our knowledge, this is the first time that CNNs have been applied directly to raw dual-energy X-ray imagery, in any field. To exploit dual-energy, we experiment with adapting several physics-derived approaches to material discrimination from the cargo literature, and introduce three novel variants. We hypothesise that CNNs can implicitly learn about the material characteristics of objects from the raw dual-energy images, and use this to suppress false positives. The best performing method is able to detect 95% of containers containing a single SMT, while raising 0.4% false positives on benign containers. This is a step change improvement in performance over our prior work.

Keywords: Cargo Screening, Automated Threat Detection, Dual-energy X-ray, Material Discrimination, Deep Learning, Convolutional Neural Networks

1. INTRODUCTION

In recent years the threat from Mumbai-style terrorist attacks using so-called Small Metallic Threats (SMTs)* on soft targets has increased. This is evidenced by the attacks in Paris and on the beaches of Tunisia. In such attacks, a few perpetrators, acting as a well-organised unit striking simultaneously against unprotected civilian targets in urban areas, can inflict mass fatalities and casualties. To prevent such attacks, it is necessary to make it more difficult for would-be terrorists to obtain SMTs. In countries, where SMTs are well controlled or illegal to possess if too powerful, would-be terrorists rely on creating sophisticated smuggling networks, which is expensive, or on purchasing SMTs from the existing pool i.e. the black market.

SMTs can be smuggled across borders using a number of modes, including by land, air and sea. They can be carried in cargo containers, parcels, passenger baggage, or independent vessels. Border agencies provide surveillance on each of these modes, and often capture X-ray images which are inspected by a human operator. However, with the vast and growing volumes of legitimate items crossing borders, reliance on human operators can slow down the throughput of items. A thriving economy relies on the unfettered movements of goods and people. Thus, there is good reason to automate elements of inspection.

Further author information: (Send correspondence to Lewis D. Griffin)
Lewis D. Griffin: E-mail: L.Griffin@cs.ucl.ac.uk, Telephone: +44 20 3108 7107

*We use the term 'small metallic threats' as we do not wish to make our research results easily discoverable by malicious actors through keyword searching. However, the smallest of the threats in question are similar in form to hand drills, whilst the largest are similar in length to a garden spade.

In recent years, there has been a flurry of activity in automated inspection of X-ray images of airport baggage[1–5] and cargo.[6–11] Of these, cargo offers the most challenge for Automated Threat Detection (ATD). Images are typically much larger, and threats easier to conceal amongst dense or complex cargo. Threats are also imaged at much lower resolution, and occupy a very small part of the image. Figure 1 shows a comparison between an RGB baggage image and a single-energy X-ray cargo image.

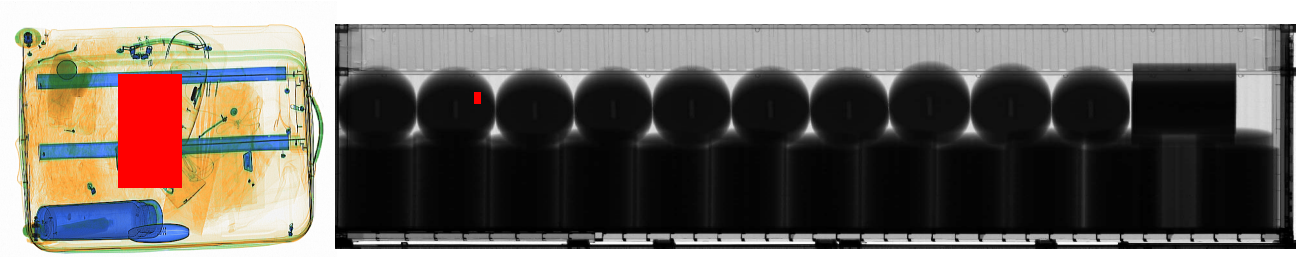

Figure 1: Comparison of a typical RGB X-ray baggage image and a typical single-energy X-ray cargo image. Each contains the same type of SMT masked by a red box, which allows comparison of SMT dimensions to the dimensions of the image.

In our previous work on cargo, we showed that using a trained-from-scratch Convolutional Neural Network (CNN), we could detect 90% of SMTs synthetically concealed in stream-of-commerce images of ISO containers, whilst raising 6% false alarms.[7] In addition, we found that by feeding the CNN the log-image as an additional input channel, the detection performance of the network was improved considerably.

In this contribution, we make two improvements on SMT detection: (i) we improve the training method used by synthesising 'spot-the-difference' concealments on-the-fly, and (ii) we investigate the use of dual-energy measurements to suppress false alarms. We seek to exploit the extra material information encoded in dual-energy images. In commercial viewing software, the raw dual-energy images can be transformed into an RGB image, useful for operators searching for threats. Instead of operating on these derived RGB images, we feed in the original dual-energy images, from which they are derived, as separate channels to the CNN. We hypothesise that the CNN can implicitly learn the types of material present in the image which can be useful for suppressing false alarms.

In the next section we review related work on ATD on X-ray imagery and dual-energy material discrimination. In Section 3, we discuss the methods used, including dataset augmentation, a new training scheme, CNN network architectures, and our approaches to utilising dual-energy information. In Section 4, we give a performance evaluation of the new training scheme and each of our dual-energy approaches, as well as examples of correctly classified and misclassified images. We conclude in Section 5.

2. RELATED WORK

We now summarise the related work on ATD and dual-energy material discrimination for cargo.

2.1 Automated Threat Detection

Most work on ATD in baggage and cargo uses Bag of Words (BoW) approaches, with a few recent publications on the use of trained-from-scratch and pre-trained deep CNNs.[2,8] Jaccard et al.[6] employ a BoW approach, oriented Basic Image Features (OBIFs)[12] histograms classified using a Random Forest (RF), to detect concealed cars in cargo containers. The approach was able to detect 100% of cargo containers containing a car, with 0.41% false positive rate. The same authors were able to improve on this work using a trained-from-scratch CNN, which resulted in a false positive rate of 0.22%.[8] This was also a 5-fold improvement on another BoW approach, Pyramid Histogram Of visual Words (PHOW).

BoW approaches have been applied to dual-energy baggage imagery, both on the raw images and the RGB material discrimination image commonly seen at airport checkpoints. Baştan et al.[1] show that the additional information from dual-energy images significantly improves object recognition. The authors, experiment with

dual-energy variants of the Scale Invariant Feature Transform (SIFT) and intensity domain SPIN image (SPIN) descriptors. They compute Colour SIFT (CSIFT) and Colour SPIN (CSPIN) descriptors, which operate on the individual colour channels of the RGB image. In addition, they compute Energy SPIN (ESPIN) descriptors which are computed directly on the raw high and low energy images. They found both ESPIN and CSPIN performed better than using SIFT or SPIN alone, with CSPIN achieving best performance.

Recently, Akçay et al.[2] have applied deep CNNs to ATD in dual-energy RGB baggage imagery. They recognise that there is a problem with training CNNs from scratch due to the limited availability of data. Thus they adopt a transfer learning approach by taking a CNN, pre-trained for general image classification tasks (ImageNet[13]), and then fine-tune it for ATD in X-ray baggage. The pre-trained CNN follows the architecture introduced by Krizhevsky et al.,[14] consisting of 5 convolutional layers and 3 fully-connected layers. The authors re-use the generalised feature extraction and representation in the lower layers of the CNN, whilst fine-tuning the upper layers. This achieves 99.26% detection and 4.08% false positives, which significantly outperforms prior work in the field.

For baggage imagery, it appears that utilising dual-energy, either by RGB image or the raw high and low energy images, can boost ATD performance. To the best of our knowledge, there have been no prior publications on ATD using dual-energy cargo images, or using CNNs directly on the raw high and low energy X-ray images, in any field.

2.2 Material Discrimination

Material discrimination in cargo works by performing image measurements at two different energies. Based on these two images, simple features can be computed such that different material atomic numbers (Z) are partitioned in feature space. These methods operate on the radioscopic transparency, T. Given the equation of image formation

$$I = \int I_0(E) e^{-\mu(E,Z)\tau} dE, \tag{1}$$

the transparency is defined as

$$T = \frac{\int I_0(E) e^{-\mu(E,Z)\tau} dE}{\int I_0(E) dE}. \tag{2}$$

In these two equations E is the photon energy, $I_0(E)$ is the initial intensity of photons emitted from the source, $\mu(E,Z)$ is the attenuation co-efficient for a material with atomic number Z for impinging photons with energy E, and τ is the material thickness. In the rest of this work we refer to the high energy and low energy transparencies as H and L, respectively.

In their seminal work for material discrimination in cargo, Ogorodnikov and Petrunin[15] use the log-ratio R and $1/H$ as features. The log-ratio is defined as

$$R = \frac{\log(H)}{\log(L)}. \tag{3}$$

This is motivated, in part, by the observation that for a monochromatic beam of energy E_γ such that

$$I_0(E) \propto \delta(E - E_\gamma), \tag{4}$$

where $\delta(\_)$ is the Dirac delta function, then the ratio of logs becomes

$$R \propto \frac{\log(e^{-\mu(E_H,Z)\tau})}{\log(e^{-\mu(E_L,Z)\tau})} = \frac{\mu(E_H,Z)}{\mu(E_L,Z)}. \tag{5}$$

So for a given material with atomic number Z, R is unique to that material and does not depend on the thickness τ. Therefore, R can be used to discriminate materials. However, for cargo screening, the X-ray photons are generated by the Bremsstrahlung process, and thus the high and low energy beams are not monochromatic, but a continuous spectra up to some cut-off energy equivalent to the energy of the accelerated electrons used to

generate the X-rays. In this polychromatic regime, material discrimination is still possible, but the log-ratio R is not unique for a particular material.

Analytic material curves in the Ogorodnikov and Petrunin[15] feature space are given in Figures 2(c&f) for monochromatic and polychromatic cases, respectively. No noise has been modelled in these examples. In the monochromatic case, the curves are well separated in feature space, however, in the polychromatic case, they overlap for small $1/H$ values (as materials become thin). The material curves are also bunched closer together, meaning that noise in the imaging system can more easily lead to material misclassification. Ogorodnikov and Petrunin[15] note that material discrimination is also difficult in the case of thick materials since noise begins to dominate the image.

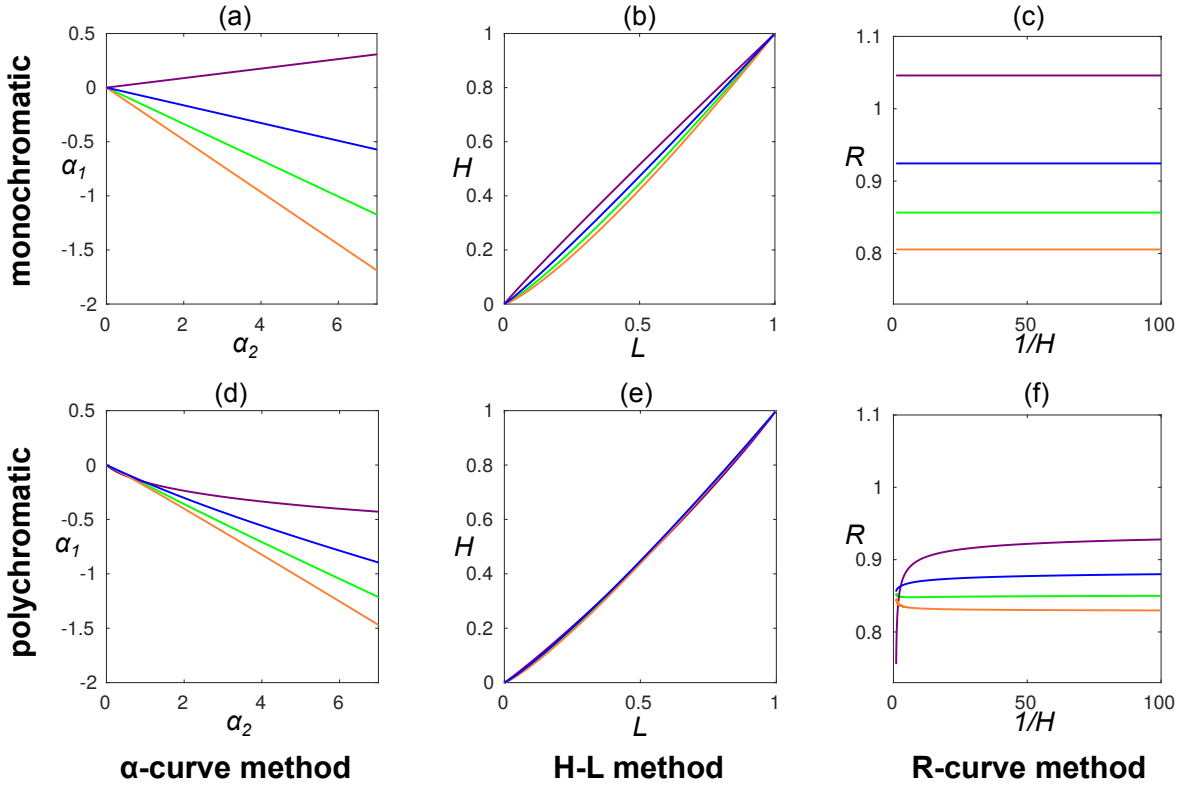

Figure 2: Material curves in feature space for three different material discrimination methods in the literature. The materials, include: Boron (Z=5, orange); Aluminium (Z=13, green); Iron (Z=26, blue); and lead (Z=82, purple). Curves were computed analytically using Equation (1), assuming a noiseless imaging system, and using attenuation coefficients from the NIST XCOM database.[16]

There are two other methods of computing features in the literature. These are known as the α-curve[17,18] and $H-L$ curve[19] methods. The α-curve method computes the features

$$\alpha_1 = -\log(H) \quad (6)$$

$$\alpha_2 = -\log(H) + \log(L), \quad (7)$$

and the $H-L$ curve method simply uses H and L as the features. Analytic material curves for these approaches are also given in Figure 2. Note that in the polychromatic case the $H-L$ curves are bunched very close together compared to the α-curve and R-curve methods, and so one would expect this method not to perform as well.

To form an RGB image, authors typically perform a system calibration by scanning known materials of varying thickness. The calibration image can be use to determine how feature space should be partitioned such that pixels can be classified into material Z groups. Each group is assigned a different hue in the coloured

X-ray image. Ogorodnikov and Petrunin,[15] found that system noise leads to a noisy RGB image and spatial information is required to improve the quality of the RGB image. The authors apply a simple segmentation algorithm and each segment is labelled as the average material over the pixels in that segment. Since the original work of Ogorodnikov and Petrunin,[15] which used a controlled laboratory set-up, other researchers have failed to replicate their accurate results for commercial systems. Some authors have focussed on detecting only high-Z materials, which can be indicative of nuclear material or adversarial shielding, citing that multi-class material discrimination is infeasible due to the levels of noise in commercial systems.[20]

In a recent review paper,[21] we argued that machine learning approaches, and in particular deep CNNs, could learn to perform material discrimination based on dual-energy measurements and spatial information in the image. However, a major hurdle to achieving this is the difficulty obtaining large datasets of different materials with pixel-level labelling. In this work we aim to implicitly learn material discrimination in order to boost performance in ATD.

3. METHODS

3.1 Dual-Energy Network Architectures

We investigate the use of different feature spaces used for material discrimination from the literature (Section 2.2) and three novel variants.

We refer to the first variant as the $\Sigma{-}\Delta$ curve method. This is similar to the $H{-}L$ curve method,[19] but rather than using H and L as features, we use $\Sigma{=}H{+}L$ and $\Delta{=}H{-}L$. This approach yields a larger separation between material curves in the feature space, and so one would expect better material discrimination as a result. The material curves for the $\Sigma{-}\Delta$ curve method are given in Figures 3(b&e).

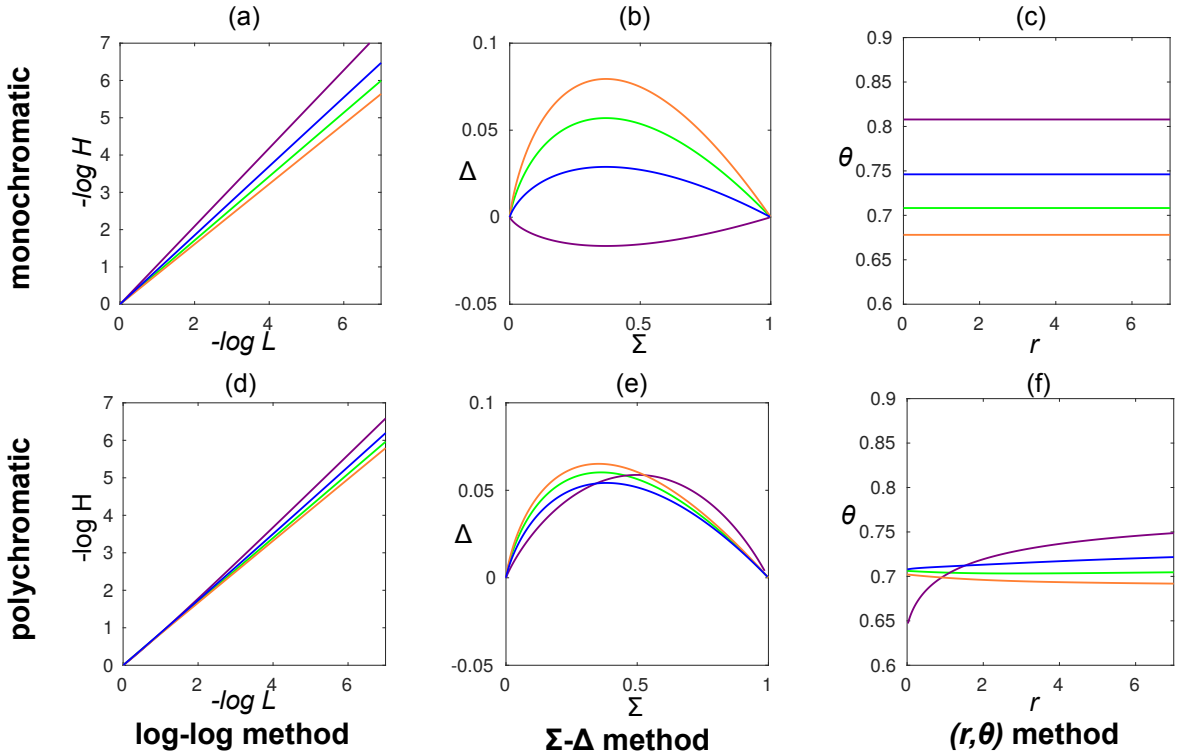

Figure 3: Material curves in feature space for three novel variants of material discrimination methods in the literature. The materials, include: Boron (Z=5, orange); Aluminium (Z=13, green); Iron (Z=26, blue); and lead (Z=82, purple). Curves were computed analytically using Equation (1), assuming a noiseless imaging system, and using attenuation coefficients from the NIST XCOM database.[16]

The second variant is the *log-log* method, which is similar to the α-curve method, but uses just $-\log H$ and $-\log L$ as features. This is convenient since, when combined with the $H-L$ curve method, it is a straightforward dual-energy generalisation, i.e.

$$\{H, -\log H\} \rightarrow \{H, -\log H, L, -\log L\}, \tag{8}$$

of our previous system, which uses H and $-\log H$ as CNN input channels.[7] This provides a good baseline for dual-energy networks.

The third variant is the (r, θ) method. This is derived by transforming the *log-log* method into polar co-ordinates:

$$r = \sqrt{(\log L)^2 + (\log H)^2} \tag{9}$$

$$\theta = \arctan(\log L / \log H). \tag{10}$$

In the monochromatic case, this give θ-values that are constant as a function of r. In the polychromatic case, the curves appear similar to the R-curve method, but there is a better separation between materials at small r-values.

As in our prior work,[7] we use CNN architectures based on the 19-layer very deep networks first described by Simonyan and Zisserman.[22] Our main modification to their original architecture, is to replace *max pooling* layers by *batch normalisation* layers.[23] Batch normalisation, which fixes the mean and variance of input distributions at each layer, is a form of network regularisation and has become preferable over *max pooling* for many researchers because it can improve training speed and ultimately performance. In total, the networks contain 16 convolutional layers, 3 fully-connected layers, and a *softmax* layer to obtain the confidence that a patch contains an SMT. The input channels are of dimension 256×256.

We investigate networks that have (i) two dual-energy input channels, and (ii) four dual-energy input channels. We have found previously that using separate input streams rather than channels, does not improve performance.[7] Illustrations of these network architectures are given in Figure 4. We begin by using two-channel networks to assess the different material discrimination methods. For each input, the high and low energy images are transformed into the feature space of a given method, and we assess performance across all methods. We next experiment with combining the best performing two two-channel inputs to construct a four-channel network, and test whether it offers improved performance.

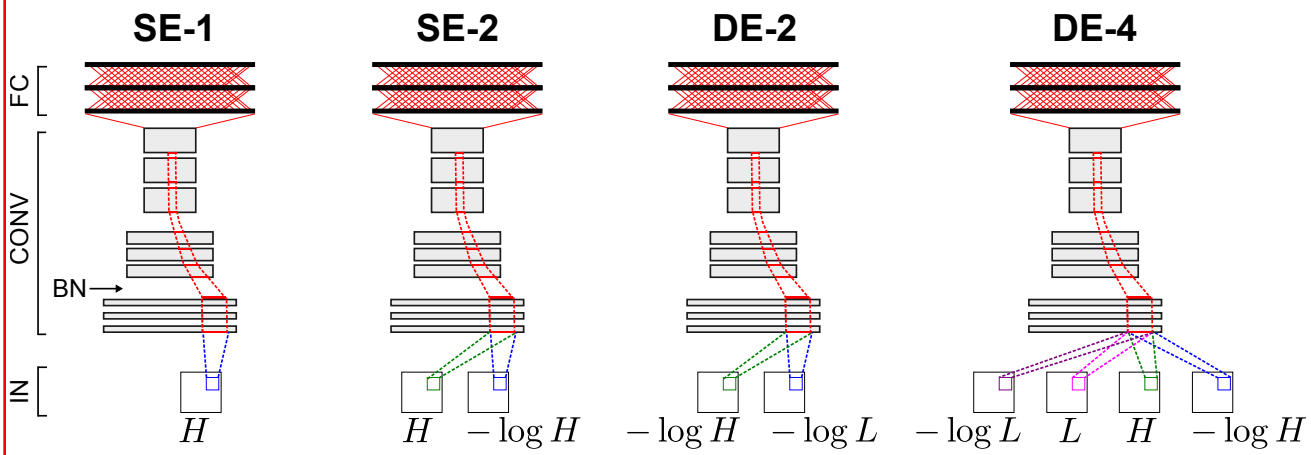

Figure 4: Illustration of the single-energy one-channel (SE-1) and two-channel (SE-2) CNN architectures used in the prior work[7] and operating on single-energy inputs (IN), and the two-channel (DE-2) and four-channel (DE-4) architectures employed in this work. Each network has a number of convolutional (CONV) layers interspersed by batch normalisation (BN) layers. The final convolutional layer feeds into the fully-connected (FC) layers. The networks used in this work consist of 16 convolutional layers and 3 fully-connected layers (not depicted in this figure).

In all experiments the CNN hyper-parameters weight decay and momentum are fixed at 10^{-4} and 0.9, respectively. In most experiments, the learning rate was decreased from 10^{-3} to 10^{-6} over the course of 30 epochs, this was reduced (and number of epochs increased) in experiments, where training was erratic. The mean image computed across the training set was subtracted from each input image.

3.2 Data, Pre-Processing and Augmentation

The benign stream-of-commerce images used in this work were acquired using a high-speed (60 km/h) Rapiscan Eagle®R60 rail scanner. The R60 is a dual-energy system which fires interlaced high and low energy beams with 4 and 6 MeV energy cut-off, respectively. Each image is 16-bit, greyscale, and ranges between 1920×850 and 2570×850 pixels for 20 ft and 40 ft long cargo containers, respectively. The pixel size is 5.6 mm, and the system has effective spatial resolution is of the order of a few mm. The image dataset contains a very diverse range of cargoes, including, but not limited to, pallets of commercial cargo, heavy machinery and industrial equipment, household goods, and bulk materials.

The total benign dataset consists of 120,000 stream-of-commerce images. Of these image we used the following dataset splits:

- 10,000 full stream-of-commerce images were reserved for testing. Of these 5,000 were kept for the benign class, and a single SMT was projected into the other 5,000 as the threat class;
- From the remaining 110,000 images, we sample 640,000 256×256 patches for training. In each CNN training batch a number of these backgrounds are randomly sampled and used to create a 'spot-the-difference' threat and benign pair (Section 3.3). A small sample of patches is also reserved for computing the validation error when training the CNN, and this is kept disjoint from the training patches.

The SMT images were acquired separately. In total, 700 SMT images were collected for a variety of different poses, types and models. The SMT instances were extracted[7] from the original scans to form a dual-energy Threat Image Projection (TIP) library. To generate *de-novo* training examples, an SMT was randomly selected from the TIP library and projected into the background patch.[24] Projecting the same SMT instance into different images results in vastly different appearances due to the translucency property of X-ray images and has recently been shown to be indistinguishable from real threat imagery.[24] The dataset is made more diverse by the injection of realistic variations including translations, intensity scaling and flipping. SMT instances were kept disjoint between the training, validation, and testing datasets.

Images were preprocessed according to our previous work:[11,25,26] (i) black columns produced by faulty detectors or source misfire were removed; (ii) source intensity variations were corrected by normalisation based on air intensity values; and (iii) salt-and-pepper pixels were replaced by the local median intensity. The high and low energy images have a small, systematic, relative translation and so to register them, translation and cropping was performed on each image pair.

3.3 New Training Scheme

We have made two modifications to our CNN training routine. First, we perform TIP and data augmentation on-the-fly, meaning that even if background patches are reused, each instantiation will use TIP with a different SMT projected under different random conditions. This means that there is a lot more variation in the threat class used in training. Second, we make the threat and benign training sets 'spot-the-difference' examples. In each training batch, each example in the threat class is identical to one in the benign class except for the projected SMT. We propose that this improves CNN training since the network can quickly learn to focus efforts on learning features for SMTs since the only discriminator between the two classes is the presence of the, often heavily shielded, SMT.

3.4 Performance Evaluation

We evaluate the performance of the different systems on full-sized container images with a single SMT projected into the container or cargo. We adopt a sliding window approach to analyse the whole image. Windows of size 256×256 pixels are sampled with a stride of 64 pixels in both the vertical and horizontal direction. For each window a confidence is computed according to the output from the *softmax* layer. A confidence score for the whole image is computed by taking the maximum of the window confidences.

We assess performance in terms of the Area Under the Curve (AUC), H-measure, and false positive rates for fixed detection rates of 90% (FPR90), 95% (FPR95) and 99% (FPR99). H-measure is a variant of the AUC that addresses issues related to underlying cost functions.[27,28]

To understand the full-image classification results we compute a heatmap of window confidences. This allows one to locate false positive or detection signals, and is potentially a useful visualisation for operators to quickly identify threats. However, since the heatmaps are computed by sliding a 256×256 window across the image, the resolution of the heatmap is poor. It would be beneficial to have a method of localising the SMT signal within this region.

To this end, we have implemented a method used by Zeiler and Fergus[29] for determining the strongest cues in a window that the CNN has used to detect SMTs. This works, by sliding a small occluding window across the 256×256 region. For each position of the occluder, the CNN score is computed. If the occluder is blocking a part of the image that provides very strong cues to the CNN, then this results in a much lower score. Thus a heatmap can be constructed which has low scores corresponding to the most SMT-like image parts, and in this way the detected SMT can be localised. This is particularly useful for the smallest SMTs when hidden amongst complicated background structure, or localising false positive detections.

4. RESULTS

Table 1 shows a summary of results for the experiments performed in this work. We discuss each in the following sections.

Arch.	Inputs/method	OtFStD?	AUC/%	H-measure	FPR90/%	FPR95/%	FPR99/%
SE-1	$\{H\}$	N	89.0	0.530	47.0	–	–
	$\{-\log H\}$	N	96.0	0.750	9.00	–	–
SE-2	$\{-\log H, H\}$	N	97.0	0.780	6.00	–	–
	$\{-\log H, H\}$	Y	98.5	0.845	1.92	7.76	34.4
DE-2	$\{H, L\}$	Y	92.2	0.624	34.3	52.2	68.5
	$\{\alpha_1, \alpha_2\}$	Y	**99.5**	0.935	0.08	0.36	15.8
	$\{R, 1/H\}$	Y	–	–	–	–	–
	$\{\Sigma, \Delta\}$	Y	96.8	0.772	7.62	24.7	56.0
	$\{-\log H, -\log L\}$	Y	98.8	0.885	1.04	3.36	19.2
	$\{r, \theta\}$	Y	–	–	–	–	–
DE-4	$\{-\log H, H, -\log L, L\}$	Y	99.2	0.900	0.56	2.38	19.4
	$\{-\log H, \Delta, -\log L, \Sigma\}$	Y	**99.5**	**0.936**	0.08	**0.34**	**15.7**
	$\{\alpha_1, \alpha_2, \Delta, \Sigma\}$	Y	**99.5**	0.922	**0.06**	0.86	18.4

Table 1: Quantitative results for all experiments. SE-N indicates single-energy architecture with N input channels, DE-N indicates dual-energy architecture with N input channels. The OtFStD (On-the-Fly Spot-the-Difference) column indicates whether the new training scheme was employed.

4.1 New Training Scheme

First we test the new training scheme, where training is done on-the-fly and the network is given 'spot-the-difference' examples in each training batch. We use exactly the same network architecture and hyper-parameters as in our previous work, and use single-energy $\{-\log H, H\}$ as input channels. The new training scheme lead to a 1.5% improvement in AUC and >4% improvement in FPR90 compared to the off-the-fly training used in our prior work.[7]

4.2 Two-Channel Dual-Energy Networks

Next we investigate how each material discrimination method performs when used alone in a two-channel network. We use the same on-the-fly 'spot-the-difference' training scheme. The $H-L$ curve method, as expected from observing the material curves in Figure 2, does not perform well in comparison to the other material discrimination methods. However, there is ∼13% improvement on its analogue single-energy network operating only on H. Thus, the $H-L$ curve method of material discrimination does provide some benefit. Similarly, the *log-log* method, yields a 2.8% improvement in AUC and 9-fold improvement in FPR90 over its single-energy analogue (operating on $-\log H$).

The $\Sigma-\Delta$ novel variant offers a significant improvement on the $H-L$ curve method, boosting the AUC by >4% and giving a 4.5-fold improvement on FPR90. This is expected from comparing the material curves in Figures 2(e) and 3(e) since the latter has better separated materials in feature space.

The best performing material discrimination method was the α-curve method. It significantly outperformed all of the other methods, yielding an AUC of 99.5% and FFPR95 of 0.36%. Again, this can be rationalised by looking at the material curves in Figures 2 and 3; the α-curve seems to offer the best separation of materials.

We found that both the R-curve and (r, θ) method were difficult to implement such that the CNN training converged, so results are unavailable. This is likely a result of the ratio of logarithms involved in both methods, and the difficulty to find a sensible range to constrain the image pixels to. Future work will focus on addressing this issue.

4.3 Four-Channel Dual-Energy Networks

The baseline four-channel dual-energy method, with inputs $\{-\log H, H, -\log L, L\}$, performs significantly better than its single-energy analogue with inputs $\{-\log H, H\}$. The AUC is improved from 98.5% to 99.2% and there is approximately a 3.5-fold improvement in FPR90. However, the method does not perform as well as the two-channel α-curve method. And so, even when working in combination, the $H-L$ curve and *log-log* methods do not perform better than the α-curve method. When swapping the $\{H, L\}$ channels with $\{\Sigma, \Delta\}$ channels, there is a large reduction in false positives, and the network now slightly outperforms the two-channel α-curve according to most metrics. Surprisingly, combining the α-curve method with $\{\Sigma, \Delta\}$ to form four-channels $\{\alpha_1, \alpha_2, \Sigma, \Delta\}$, performance worsens. We are unclear of the reason for this.

Out of all the methods tested, the four-channel network with $\{-\log H, \Delta, -\log L, \Sigma\}$ inputs performs best across most performance metrics. However, it is worth noting the average image processing times of the networks. The two-channel α-curve method takes an average of 2.45 s to process an image, whereas the four-channel network takes an average of 3.05 s. In an operational context, the former can process an extra 25% cargo containers over a given time period, or require 25% less computing resources. So the α-curve method may be the more attractive solution in practise, since it only leads to a very small number of extra false positives.

4.4 Classification Examples

In this section we provide some example classification results for the best four-channel network, which combines the *log-log* and $\Sigma-\Delta$ methods. We set the detection threshold to give a 95% detection rate. In Figure 5, we give example detections. Some of these examples are very difficult for even humans to detect in a zoomed-view, so localisation heatmaps[29] are used to indicate where the CNN is picking up strong SMT cues. The strongest cues tend to be on the part of the SMT which is most visible in the image. This is particularly noticeable in Figure 5(b&c), where the majority of the SMT is shielded by very dense cargo, and the strongest cues are located on the small parts of the SMT that are unshielded. In Figure 5(a&e), the SMTs are densely shielded, however they are still visible in the log-transformed image. In Figure 5(d), the SMT has been concealed on complicated background cargo, which makes it difficult to locate by eye.

Figure 6 shows an example of a false positive. The strongest cues are located at the junction between the floor and two parallel metal tubes. In this example, the false positive appears locally as SMT-like and is also made out of a similar material. In Figure 7, we give examples of false negatives together with the CNN inputs for the SMT patch. In both cases the SMT cannot be seen in the inputs and thus there is very little, if no, information for the network to work on.

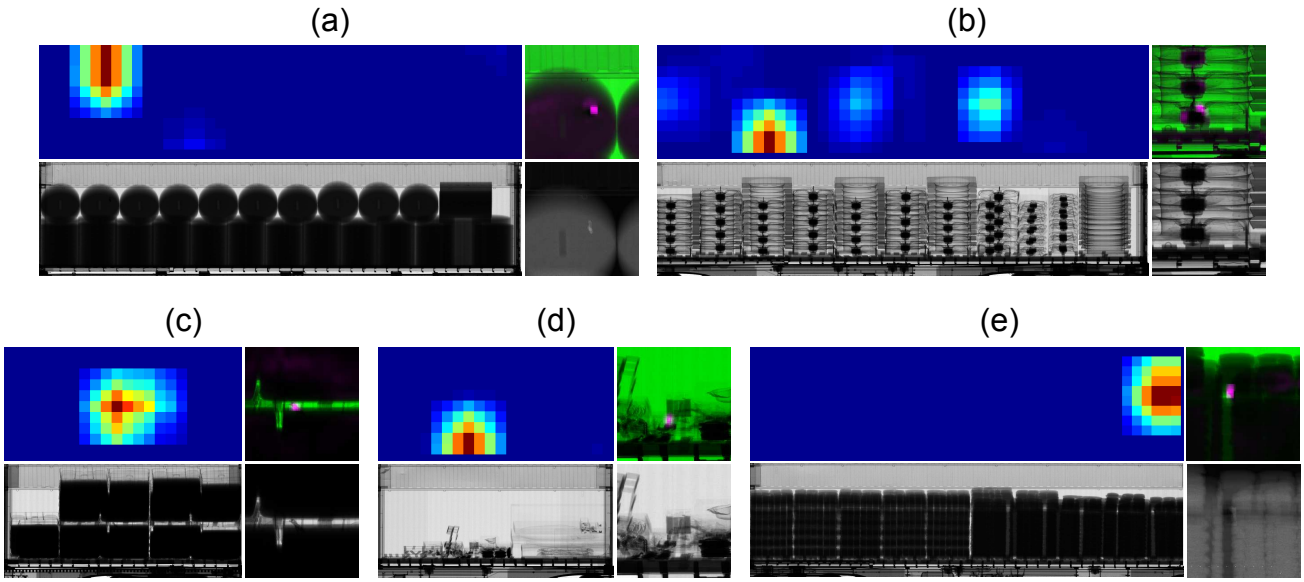

Figure 5: Examples of firearm detections in full cargo containers. For each example, going clockwise from top left: (i) score heatmap - red indicates high confidence of firearm; (ii) localised heatmap overlaid on raw image patch - pink indicates strong SMT cues picked up by CNN; (iii) raw or log-transformed image patch depending on which is easiest to see SMT; (iv) original raw image.

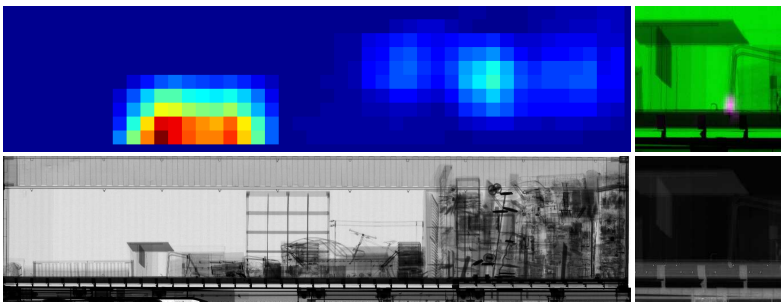

Figure 6: Example false positive. Going clockwise from top left: (i) score heatmap - red indicates high confidence of firearm; (ii) localised heatmap overlaid on raw image patch - pink indicates strongest SMT cues picked up by CNN; (iii) log-transformed image patch; (iv) original raw image.

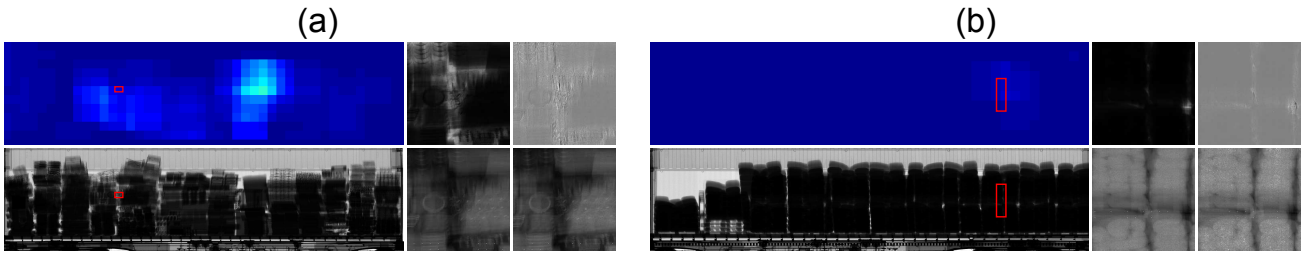

Figure 7: Example false negatives. Above each raw image, we give the heatmap of the individual window scores. To the right of each, we give the four patches that are inputs to the network. The red rectangles indicate the groundtruth location of the SMT.

5. CONCLUSION

We have investigated the use of trained-from-scratch Convolutional Neural Networks (CNNs) on complex dual-energy X-ray cargo imagery. Dual-energy X-ray imagery can, in theory, reveal information about the types of material in the image. Coupled with spatial information, we hypothesise that CNNs can implicitly learn to exploit the material information to suppress the number of false alarms in Automated Threat Detection (ATD). We have built upon our prior work,[7] on ATD of so-called Small Metallic Threats (SMTs), by investigating three existing methods of cargo material discrimination[15,17–19] and three novel variants, and how they perform when fed to the CNN as separate input channels. We have also introduced a new training scheme, which creates 'spot-the-difference' threat and benign training pairs, complete with data augmentation, on-the-fly.

In this investigation, we found that each dual-energy CNN performed better than its single-energy analogue. This supports our hypothesis that the CNN can implicitly learn how to decode material information from dual-energy images in order to suppress false alarms. Moreover, we found that a four-channel input, consisting of the sum and difference of the high and low images together with the log-transform of the high and low energy images, yielded the best performance in terms of the number of false positives given fixed detection rates of 95% and 99%. This system was capable of detecting 95% of containers containing a single SMT, while raising 0.34% false positives on benign containers. Overall, our improvement by exploiting dual-energy and the new training scheme, resulted in a 2.5% improvement in Area Under the Curve (AUC), and a 100-fold improvement in the false positive rate when the detection rate is fixed at 90%. Another way of expressing this improvement, is that if we kept the false alarm rate at the level of our prior work (6%), we can now detect 98.4% of SMT; an improvement of 8.4% in detection.

Furthermore, we have identified that in an operational context, it might be preferable to use a two-channel CNN based on the α-curve[17,18] method of material discrimination. This is because, whilst its performance is comparable to the best four-channel network, it is capable of processing full images at a faster rate; it would allow 25% more cargo containers to be inspected in a given time period.

Acknowledgement

The datasets used in this work were provided Rapiscan Systems.

REFERENCES

[1] Baştan, M., Byeon, W., and Breuel, T. M., "Object Recognition in Multi-View Dual Energy X-ray Images.," in [*BMVC*], (2013).

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

[3] Kundegorski, M., Akcay, S., Devereux, M., Mouton, A., and Breckon, T., "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*], IET (November 2016).

[4] Flitton, G., Mouton, A., and Breckon, T. P., "Object classification in 3D baggage security computed tomography imagery using visual codebooks," *Pattern Recognition* **48**(8), 2489–2499 (2015).

[5] Baştan, M., Yousefi, M. R., and Breuel, T. M., "Visual words on baggage X-ray images," in [*Computer analysis of images and patterns*], 360–368, Springer (2011).

[6] Jaccard, N., Rogers, T. W., and Griffin, L. D., "Automated detection of cars in transmission x-ray images of freight containers," in [*Advanced Video and Signal Based Surveillance*], 387–392, IEEE (2014).

[7] Jaccard, N., Rogers, T. W., Morton, E. J., and Griffin, L. D., "Automated detection of smuggled high-risk security threats using Deep Learning," *arXiv preprint arXiv:1609.02805* (2016).

[8] Jaccard, N., Rogers, T. W., Morton, E. J., and Griffin, L. D., "Detection of concealed cars in complex cargo X-ray imagery using Deep Learning," *Journal of X-Ray Science and Technology* (Preprint), 1–17 (2016).

[9] Andrews, J. T., Morton, E. J., and Griffin, L. D., "Detecting anomalous data using auto-encoders," *International Journal of Machine Learning and Computing* **6**(1), 21 (2016).

[10] Zhang, J., Zhang, L., Zhao, Z., Liu, Y., Gu, J., Li, Q., and Zhang, D., "Joint shape and texture based X-ray cargo image classification," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*], 266–273 (2014).

[11] Rogers, T. W., Jaccard, N., Morton, E. J., and Griffin, L. D., "Detection of cargo container loads from X-ray images," in [*The IET Conference on Intelligent Signal Processing*], (2015).

[12] Griffin, L. D., Lillholm, M., Crosier, M., and van Sande, J., "Basic image features (bifs) arising from approximate symmetry type," in [*International Conference on Scale Space and Variational Methods in Computer Vision*], 343–355, Springer (2009).

[13] Deng, J., Dong, W., Socher, R., jia Li, L., Li, K., and Fei-fei, L., "Imagenet: A large-scale hierarchical image database," in [*Conference on Computer Vision and Pattern Recognition*], (2009).

[14] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [*Advances in neural information processing systems*], 1097–1105 (2012).

[15] Ogorodnikov, S. and Petrunin, V., "Processing of interlaced images in 4–10 MeV dual energy customs system for material recognition," *Physical Review Special Topics-Accelerators and Beams* **5**(10), 104701 (2002).

[16] National Institute of Standards and Technology, "NIST XCOM: Photon Cross Sections Database," (v1.5) (2017).

[17] Li, L., Li, R., Zhang, S., Zhao, T., and Chen, Z., "A dynamic material discrimination algorithm for dual mv energy x-ray digital radiography," *Applied Radiation and Isotopes* **114**, 188–195 (2016).

[18] Novikov, V., Ogorodnikov, S., and Petrunin, V., "Dual energy method of material recognition in high energy introscopy systems," *Questions of Atomic Science and Technology [translated from Russian]* (1999).

[19] Zhang, G., Zhang, L., and Chen, Z., "An HL curve method for material discrimination of dual energy X-ray inspection systems," in [*Nuclear Science Symposium Conference Record*], **1**, 326–328, IEEE (2005).

[20] Fu, K., Ranta, D., Guest, C., and Das, P., "The application of wavelet denoising in material discrimination system," in [*IS&T/SPIE Electronic Imaging*], 75380Z–75380Z, International Society for Optics and Photonics (2010).

[21] Rogers, T. W., Jaccard, N., Morton, E. J., and Griffin, L. D., "Automated X-ray image analysis for cargo security: Critical review and future promise," *Journal of X-Ray Science and Technology* **25**(1), 33–56 (2016).

[22] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[23] Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167* (2015).

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

[25] Rogers, T. W., Ollier, J., Morton, E. J., and Griffin, L. D., "Reduction of wobble artefacts in images from mobile transmission x-ray vehicle scanners," in [*International Conference on Imaging Systems and Techniques*], 356–360, IEEE (2014).

[26] Rogers, T. W., Ollier, J., Morton, E. J., and Griffin, L. D., "Measuring and correcting wobble in large-scale transmission radiography," *Journal of X-Ray Science and Technology* **25**(1), 57–77 (2016).

[27] Hand, D. J., "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine learning* **77**(1), 103–123 (2009).

[28] Hand, D. J. and Anagnostopoulos, C., "A better Beta for the H measure of classification performance," *Pattern Recognition Letters* **40**, 41–46 (2014).

[29] Zeiler, M. D. and Fergus, R., "Visualizing and understanding convolutional networks," in [*European conference on computer vision*], 818–833, Springer (2014).