"Unexpected Item in the Bagging Area": Anomaly Detection in X-Ray Security Images

Lewis D. Griffin, Matthew Caldwell, Jerone T. A. Andrews, and Helene Bohler

Abstract—**The role of anomaly detection in X-ray security imaging, as a supplement to targeted threat detection, is described, and a taxonomy of anomaly types in this domain is presented. Algorithms are described for detecting appearance anomalies of shape, texture, and density, and semantic anomalies of object category presence. The anomalies are detected on the basis of representations extracted from a convolutional neural network pre-trained to identify object categories in photographs, from the final pooling layer for appearance anomalies, and from the logit layer for semantic anomalies. The distribution of representations in normal data is modeled using high-dimensional, full-covariance, Gaussians, and anomalies are scored according to their likelihood relative to those models. The algorithms are tested on X-ray parcel images using stream-of-commerce data as the normal class, and parcels with firearms present the examples of anomalies to be detected. Despite the representations being learned for photographic images and the varied contents of stream-of-commerce parcels, the system, trained on stream-of-commerce images only, is able to detect 90% of firearms as anomalies, while raising false alarms on 18% of stream-of-commerce.**

Index Terms—**Anomaly detection, object categorization, security imaging, threat detection, X-ray imaging.**

I. Introduction

"UNEXPECTED item in bagging area," the too familiar refrain of supermarket self-service checkouts [1], neatly expresses the aim of this research: to determine when an X-ray imaged bag or parcel has unusual contents.

A. Automation in X-Ray Security Imaging

X-ray imaging is used to inspect luggage, mail and vehicles to detect and discourage transport of illegal or dangerous items [2]; such as Improvised Explosive Devices (IEDs) within baggage [3], ivory within cargo [4], and firearms within parcels [5]. X-ray security scanners have become more sophisticated over recent decades, adopting multi-view systems that allow 3-D structure to be interrogated [6]; and multi-energy acquisition, allowing false coloring for material discrimination [7]. By our analysis, trained image inspectors use up to four modes of inspection, depending on the scenario:

Threat Detection (TD) - looking for specific classes of item e.g. IEDs.

Semantic Analysis (SA) - assessing broad attributes of the scanned contents, such as *illegality*, *danger* and *high value*. These attributes align with the over-arching goals of the screening process, and could be present even though the specific scanned items were not present in a watch list of threat items.

Manifest Verification (MV) - confirming that contents match a declaration. For cargo containers these are expressed in terms of HS codes [8].

Anomaly Detection (AD) - looking for deviations from normal that may indicate concealment or subterfuge.

These are complex operations, and so the inspection process remains error-prone, costly and time-consuming [9]. If they could be automated there would be benefits in cost, speed, consistency, and reduced opportunities for corruption [10]. Automated systems are typically used to reduce the number of items that require visual or manual inspection by an operator. However, since human inspectors operate several inspection modes in parallel (e.g. SA & AD while doing TD), it may be necessary to achieve automation of multiple modes, so that security effectiveness is not impacted.

Automation in the X-ray security domain area has focused on TD. Systems have been described that target a particular class of objects (e.g. cars [11], firearms [12]–[14], laptops [15]), and have demonstrated performance comparable to humans when using a convolutional neural network (CNN) [16]. Algorithms for MV are less well developed but include [17], [18]; while algorithms for SA of X-ray security images have not yet been described. As well as technical challenge, sourcing the data needed for the development, training and testing of SA and MV algorithms is clearly a major obstacle. The focus of the current work is AD, which in images has been extensively treated for satellite imagery (e.g [19]–[21]), less commonly in video (e.g. [22]–[24]), and rarely in security X-ray [25]–[28].

B. Anomaly Detection in X-Ray Security Imaging

Automated AD has been proposed as a useful function in many domains. In some applications the aim is to detect anomalies (instances, events or states) that can be considered low probability extremes of normal variation [29]; in other applications, including X-ray security, the anomalies to be detected arise when a different generating process takes over from the normal one [30], in particular when an adversary is attempting a damaging or illegal action [31].

Manuscript received March 26, 2018; revised September 22, 2018 and October 30, 2018; accepted October 31, 2018. Date of publication November 16, 2018; date of current version February 13, 2019. This work was supported by the U.K. Home Office under Grant DSTLX1000112457 and Grant DSTLS1000121403. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pedro Comesana. *(Corresponding author: Lewis D. Griffin.)*

L. D. Griffin, M. Caldwell, and J. T. A. Andrews are with the Department of Computer Science, University College London, London WC1E 6BT, U.K. (e-mail: l.griffin@cs.ucl.ac.uk).

H. Bohler was with the Department of Computer Science, University College London, London WC1E 6BT, U.K. She is now with Spacemaker, 0166 Oslo, Norway.

Digital Object Identifier 10.1109/TIFS.2018.2881700

We propose the following taxonomy of anomaly types in X-ray security. Major types:

Appearance – an unusual shape, texture or density e.g. due to an explosive in powder form [32].

Semantic – an unusual category of object. The appearance of unfamiliar objects is less well-known to image inspectors, so they afford great opportunity for concealment e.g. a recent IED concealed within a meat mincer [33].

Appearance-given-Semantics – IEDs have been concealed amongst the components of complex objects (e.g. electronic devices), or by replacing components with imitations. Done well this will not disrupt the recognisability of the object, but close inspection may reveal subtle differences from normal for that category.

Minor types:

Relative-appearance – a subset of items appearing different from the others e.g. one pallet of lemons in a cargo looking different because they conceal narcotics [34].

Arrangement – Unusual packing or voids indicating concealment.

Low-level – A malformed image indicating that rescanning is required.

Co-occurrence – An unusual collection of objects to see together.

Passenger/route-relative – Unusual to see on that flight route, or associated with that profile of passenger.

In this work we are concerned with Appearance Anomalies and Semantic Anomalies.

C. Approaches to Anomaly Detection

Any approach to AD has two main parts: a representation for the data; and a method, making use of that representation, for scoring the outlier status of a test item relative to a sample of normal data. A broad conclusion from our previous work in Anomaly Detection [26], [27] is that the choice of representation is the more critical component.

1) Data Representation: The challenge of representation in Anomaly Detection is that a meagre representation may not clearly express the features that make an anomaly unusual, while a too generous representation risks making every datum unique, and anomalies not any *more* unique. Three approaches to data representation can be distinguished: raw, engineered or learned.

Even in TD the *raw data representation* is often ineffective because the discriminative features of the data are non-linear combinations of dimensions, masked by irrelevant dimensions. Regularization methods can help with this by, for example, preferring sparse representations, but their effect needs balancing against performance, which cannot be evaluated in AD at training time.

Engineered representations can bring out the important features of the data and suppress the irrelevant. Such engineering is viable in TD, when training data can guide the process, but is well-known to be difficult, with no guaranteed recipe for success. In AD engineering an effective representation is even more difficult as little or nothing is known about the anomaly class.

Learned representations, such as computed by convolutional neural networks, are the state-of-the-art for TD, and can be extremely effective when sufficient data is available to constrain the learning. No simple equivalent of these methods is available for AD, because it is precisely the performance at predicting the training data labels that drives the learning.

In previous work we have used two methods for learning representations for AD: auto-encoders and internal labels.

We trained auto-encoder networks [35] on normal data, and used as representations the pattern of reconstruction errors of a datum and/or the hidden layer activations of the network. This had some success for detecting firearms concealed within the fabric of empty cargo containers, since the auto-encoder was able to capture the limited variability of the normal class in this case [27]. However, the auto-encoder approach has little prospect of coping with the variability of the data within bags and parcels. Our experience with firearms within empty cargo containers supports that, as does the finding that a class (e.g. 1) of MNIST digits were only detected as anomalies relative to a normal set of the other digit classes (0-9) with an average area-under the ROC curve (AUC) of 73% using anomaly detection based on a variational autoencoder [36].

As an exemplar of the internal labels approach, we subdivided a large gallery of male face photos into subsets according to the identity of the subject, and trained a network to judge whether a pair of images showed the same or different people. We then used the final layer activations of that network as a representation of face photos. Using these representations we were able to spot female faces as anomalous relative to a normal class of male faces [26]. Such an internal labels approach to learning representations for anomaly detection is very attractive but is inapplicable to the current bag/parcel problem as we do not have a rich set of labels on normal parcels to use to drive the representation learning.

The new approach we pursue in this work is related to the internal labels approach; but rather than train on normal data, for which internal labels are unavailable, we train on a related dataset for which they are, hence a type of transfer learning [26], [37]. The related dataset is photographic (i.e. non X-ray) images with semantic content labels. In fact, since semantic classification in photographs is a well-studied problem [16], [38], we have no need to assemble a dataset and train a new network, we can instead use one pre-trained on a large amount of data. We hypothesize that this approach will have some degree of success as X-ray images and photographic images have much in common – even an untrained viewer can recognize some objects in security images.

2) Outlier Detection: Methods for detecting outliers (anomalies), relative to a sample of a normal population, have been proposed based on: boundaries, trees, distances, and densities.

One-class SVM methods encircle the normal samples with a boundary beyond which a test datum is classified as anomalous [39]. Isolation Forests compute an anomaly score for a test datum as the average number of sequential threshold tests that need to be applied to separate it from the normal data; where threshold dimensions are chosen randomly, and threshold values are chosen uniformly from the range of data that has not yet been split off [40]. Distance-based measures use the mean distance to the k-nearest normal data items [41], or compare that value to the local average [42].

In density methods a test data is scored by its likelihood according to an estimate of the density of the normal population. The density estimate can be arrived at by fitting a parametric form [43], or by kernel-density-estimation (kde) [44]. Parametric fitting will only be effective if the true population has the fitted form; whereas kernel-density-estimation requires bandwidth selection, possibly spatially-varying and anisotropic, which is known to be a difficult problem [45].

Specialized methods exist for estimating the density of a population of binary vectors. If it can be assumed that the distribution is dimension-separable then a product of Bernoulli distributions (naïve Bayes) is the ideal approach, and can still be effective when the assumption is violated [46], [47]. When independence is not assumed, a variety of approaches have been suggested.

- Extensions of naïve Bayes that remove correlated dimensions give inconsistent performance [48].
- The quadratic exponential model assumes that log-likelihoods are a linear function of the dimensions of a vector and all its pairwise products [49]. In principle the weights that model a sample can be determined by maximum likelihood (ML) estimation, but the computation is impractical for large dimensional data.
- Neural networks that learn an estimate of the distribution have been proposed [50].
- Dichotomization of a multivariate Gaussian by passing each dimension of a random variate through a Heaviside function, will generate binary vector data [51]. In principle the mean and covariance of the Gaussian could be set by ML-estimation, though this seems a very difficult computation.
- Modeling the distribution as a multivariate Gaussian, ignoring its binary nature [52].

In this work we will use parametric density estimation approaches for anomaly detection, since they handle high-dimension well, and we will show that our data is close to parametric form. For semantic anomalies, the scalar-valued vector representations we use make the distribution of normal data well-modeled by a multivariate Gaussian. For appearance anomalies, the binary vector representations we use are crudely modeled by a multivariate Gaussian, but we propose an adjusted variance computation that improves the modeling.

D. The Proposed Approach

We present approaches to detection of Appearance Anomalies and to detection of Semantic Anomalies. As a test problem we use X-ray images of parcels. The normal set are UK stream-of-commerce (SoC) parcels containing diverse contents (but not firearms). The anomaly set are staged-threat (threat) parcels containing normal contents plus a firearm. We stress that we are using this as a test problem for AD, and are not suggesting that firearms are not better detected by a TD method trained on firearms data. We propose that performance at this task will give an indication of performance at detecting other objects that are absent from the SoC.

Our approach to AD makes use of representations computed by a CNN classifier trained to categorize a wide range of objects in *photographic* images. We use a representation based on the final pooling layer of the CNN for detecting appearance anomalies, and a different representation based on the final logit layer for detecting semantic anomalies. For both types of anomaly we detect outliers by using likelihoods computed according to a Gaussian model of the density of SoC data. The details of the Gaussian models are different for the two types of anomaly.

In section II we describe the image datasets. In III we describe the representations. In IV we describe TD, of firearms, using these representations. These results allow us to establish that the representations used have the potential to detect firearms as anomalies. In V we present the details of our methods for AD, and give results for appearance anomalies, semantic anomalies, and combined anomalies of either type. In VI we summarize and conclude.

II. Datasets

We use an image dataset assembled and constructed by the Centre for Applied Science & Technology (CAST), part of the Home Office of the UK Government. The data was prepared for development and testing of TD algorithms.

The data consists of X-ray images of parcels, in two sets (Fig 1). The *stream-of-commerce* (SoC) set shows 5000 parcels collected from a UK parcel distribution center. In a fraction of these images objects, such as machine and computer parts, clothing and footwear, can be recognized, but in the majority the contents are less obvious. The *staged-threat* (threat) set shows 234 parcels, each packed with benign objects, selected as usual for parcel contents, plus a firearm of pistol, carbine or rifle type, in some cases partially disassembled.

All parcel images are dual-view (i.e. a pair of images), acquired from roughly perpendicular directions, and false-colored based on dual-energy imaging. Images are 764 pixels high; SoC images have a median width of 676 pixels (IQR = [507, 906]), while threat images have a tendency to be larger, with a median width of 990 pixels (IQR = [515, 1161]). The model and make of scanner used was not disclosed by CAST for reasons of commercial neutrality, but are presumed to be tunnel scanners as seen in airports and large mailrooms. It is unclear whether the two datasets were collected using the same scanner, but nothing in the images suggests they were not.

The firearms within the threat parcels vary in how difficult they are to recognize, dependent on their size and how they lie relative to the other parcel contents. Subjective assessment by the authors categorized the firearms as being *easily seen* in 81% of parcels, *difficult to see*in 15%, and unrecognizable in 3%. The diversity of SoC parcels and the range of firearm visibilities in threat parcels is illustrated in Fig. 1.

As a pre-processing step all images were first automatically cropped to remove air around the parcel based on thresholding and connected components analysis (parameters tuned by experimentation). Next, each was reduced to an unstructured set of 224×224 pixel patches using a stride of 112 or less so that both images in a dual-view pair were uniformly

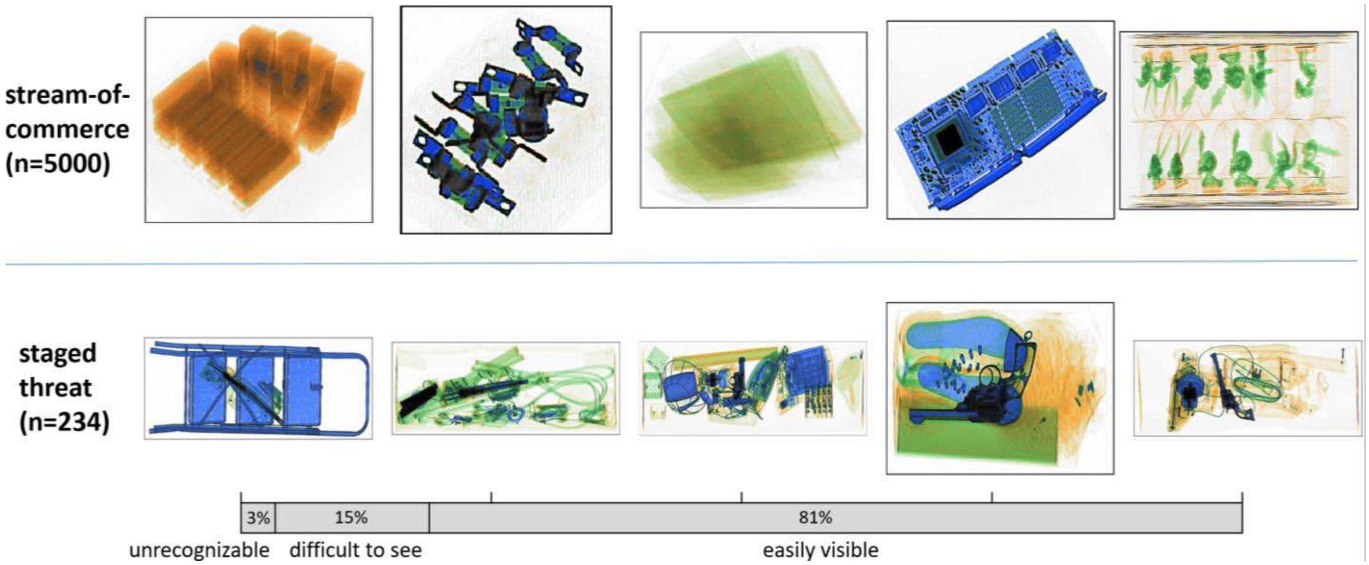


Fig. 1. Example parcel images. In all cases the more informative of the dual views is shown. In staged threats (bottom row) the firearms vary in visibility, as indicated by the grey bar along the bottom. The examples shown roughly correspond to the visibilities indicated by the tick marks above the bar.

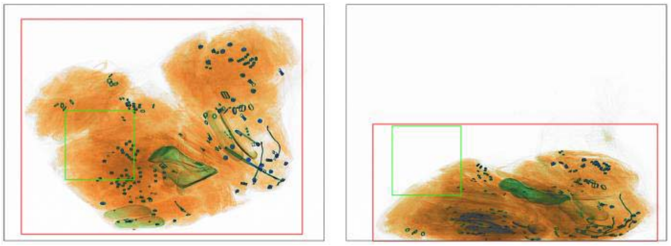


Fig. 2. Example dual-view SoC parcel image. Black rectangles show the full extent of the image. Red rectangles show the automatically identified cropping boundaries removing air around the parcel. Green squares show the size of patches that the cropped areas are reduced to. Patches overlapped by 50% or more in each dimension.

covered, extending into the corners (see Fig. 2). An average of 26 patches were produced for each dual-view SoC image.

III. Representations

We compute representations of the image appearance and semantic content using the Wolfram ImageIdentify CNN included as part of Mathematica (v11.1) [53]. We examine the effect of using alternative CNNs in section V.F. This CNN takes a 224×224 RGB image as input and produces a vector of classification confidences for 4315 semantic categories. The CNN is very similar in architecture to Inception V3 [54] but was chosen for its larger number of semantic categories. It has 232 layers, ∼15M parameters, and a trained size of 65MB. After multiple layers of convolution, pooling, batch normalization and RELU non-linearity; the activations of a final 1024-D pooling layer expressing presences of image-wide appearance features; then the activations of a 4315-D logit layer express evidence for semantic categories, computed as linear functions of the pooling layer activations; and a final softmax layer compresses the logit layer activations into a unit-sum histogram of positive confidences over the 4315 semantic categories.

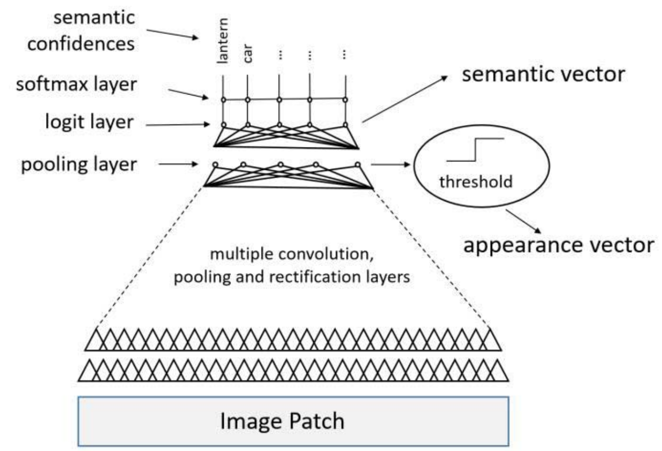


Fig. 3. Extraction of vector representations suitable for anomaly detection from a generic object identification CNN.

Figure 3 illustrates how we intercept the computations of the Wolfram Image Identify CNN to extract appearance and semantic vectors for use in anomaly detection.

A. Appearance

The appearance representation is based on the 1024 activations of the final pooling layer of the network. These activations are non-negative values that can be interpreted as signaling the degree of presence of a complex structure within the image. For CNNs in general, understanding what the responses at later layers of the network signal has proved difficult, but some at least indicate particular textures, while others particular shapes (e.g. faces) [55].

We have a couple of expectations about how appearance anomalies will manifest in the images which we use to guide the design of our processing pipeline, hopefully without biasing the anomaly detection away from generality. First is that appearance anomalies will be localized rather than

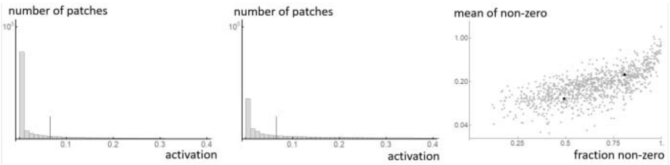


Fig. 4. Illustrates the distribution of values in the final pooling layer of the Wolfram ImageIdentify Net applied to patches from the stream-of-commerce data. Left and Centre: the distributions of values for two example dimensions; variation in the fraction of non-zero values, and their mean, is apparent. The threshold value used for binarization is indicated in the plots. Right: one plot point for each of the 1024 dimensions showing how their distributions vary; the point corresponding to the left and center histograms are marked more darkly.

diffusely present. Second is that their presence may be signaled not just by high activations in particular dimensions of the appearance representation, but by a pattern of high and low activations. These expectations motivate our choice to compute separate representations for each image patch, and maintain these as separate elements in an unstructured set; rather than, say, forming a single representation which is the maximum activation for each dimension, over the patches of an image; or by reducing the resolution of the image so that it can be input into the CNN in its entirety.

Looking at the activations in bulk, across the patches of the SoC dataset, reveals that the distributions of values in each dimension are qualitatively similar (Fig. 4 left and center). Each has a substantial fraction of zero-valued activations, with the remainder distributed approximately exponentially. They differ quantitatively though: the fraction of non-zero activations varies, as does their mean (Fig. 4 right).

The distributions of activations from the pooling layer are a hybrid of a categorical variable (zero vs. non-zero) and a continuous (if non-zero). This is very far from Gaussian, which is our preferred model for the distribution of SoC data. To make the data more approximately normal we binarize all activations by thresholding (below threshold maps to 0, above to 1). We use a common threshold for all dimensions; reasoning that they can be considered roughly commensurate as they all feed into the same linear-weighted logit functions in the next layer of the network.

To choose the binarization threshold, we consider how extreme are the binary distributions that result. More extreme distributions (i.e. mostly 0s or mostly 1s) are approximated less well by gaussians e.g. if a binarized channel is 80% 0s and 20% 1s, then a Gaussian approximation will have mean 0.2 and standard deviation 0.4; for this Gaussian, the ratio of the probabilities of a 0 and a 1 is 6.5, not much bigger than 4.0=80/20; whereas for a 90%/10% split the ratio becomes 85.1, much larger than 9.0=90/10. We find that a threshold of 0.065 (marked in Fig. 4 left and centre) best avoids extreme binary distributions at either end of the range. We evaluate the sensitivity of this choice in section V.E.

To summarize: as a representation of appearance we use a 1024-D binary vector for each image patch, computed by thresholding the final pooling layer activations from the Wolfram ImageIdentify Net. The appearance representation of a parcel is the unstructured set of representations of patches from both views.

TABLE I

GUN CATEGORIES

bullet-firing gun	assault rifle	automatic pistol
	Bren gun	carbine
	Luger	machine pistol
other gun	paintball gun	
shell-firing gun	bazooka	Bofors gun
	cannon	field artillery
	howitzer	

The 12 categories of gun among the 4315 semantic categories that WolframNet classifies into. Only the five shell-firing guns are considered parcel-implausible

B. Semantic

We base a semantic representation for each image on the 4315 activations of the logit layer of the network. Each activation expresses the evidence for a different semantic category. In the softmax layer that follows, these activations are competed against each other to sharpen the response towards the largest ones for an image, but in this layer they are independent assessments. The categories consist of 19 famous buildings, 47 fictional characters and 4249 concrete concepts. Of the concrete concepts we have manually identified 996 as being parcel-plausible, the others being too large (e.g. snowdrift), living (e.g. red snapper) or food items (e.g. hamburger). The concrete subset contains 12 categories which are a type of barrelled projectile weapon (gun for short), of which 7 are in the parcel-plausible subset (Table I).

In contrast to our approach for appearance, for the semantic representation we use a maximum-over-patches operation to form a single representation for an entire image rather than a per-patch representation. This is because we expect semantic anomalies to be manifest as larger-than-normal activations in single dimensions, and so there is no advantage in maintaining separate patch representations. We still process the image by patches, rather than downsampled as a single input, as we consider the patches are reasonably well-matched to typical object sizes in parcels.

The logit layer activations are real values which, when examined in bulk across the SoC dataset, are well approximated as Gaussian distributions in each dimension (Fig. 5). The mean and standard deviation of the distributions vary, with a tendency for standard deviation to increase with mean, which we model by a linear relation between mean and log standard deviation (Fig. 5).

For each image we have determined the category with the highest activation. Table II shows the most common categories that result. It can be seen that these bear little relation to what we expect parcels to contain; parcel-implausible ones are often selected; and while guns do register in the threat dataset they are picked much less frequently than they are in fact present. This shows that the Wolfram ImageIdentify Net applied to X-ray parcel images performs very poorly in the normal mode of usage, not surprising given how different most objects appear in X-rays compared to the photographic images on which the net was trained.

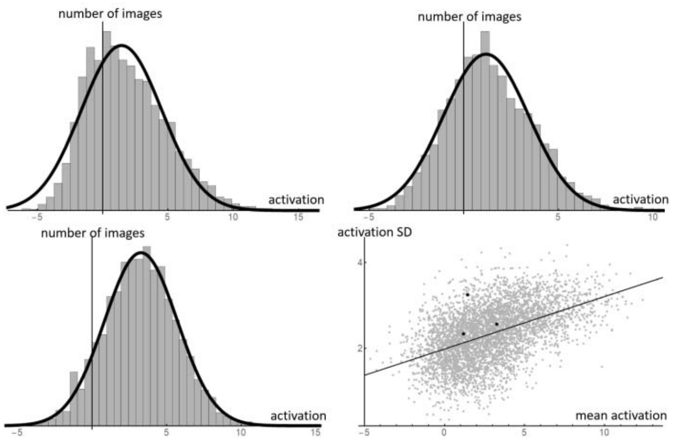


Fig. 5. Illustrates the distribution of values in the logit layer of the Wolfram ImageIdentify Net (maxed over patches in an image). Each histogram shows the distribution for a single logit layer node, which corresponds to a particular semantic category: *Colosseum*, *sacred ibis* and *sundew* in the examples shown. Logit layer values show the evidence the CNN has found for a particular category (large values = more evidence). In the next softmax layer the values are exponentiated and competed across categories to produce confidences in the range [0,1], but it is the raw logit values that are used for anomaly detection. The three example distributions were chosen to illustrate the range of closeness to Gaussian form of the distributions (respectively worst, median and best). Although the distributions are always close to Gaussian form they vary in the means and standard deviations, with some correlation between the two as shown by the bottom right plot. Note the log scale for the standard deviation in this plot.

TABLE II

COMMON DETECTED CATEGORIES

stream-of-commerce		staged-threat	
rule	7.3%	rule	5.9%
envelope	6.1%	**volleyball net**	4.2%
stratus	3.6%	envelope	3.3%
fluorescent lamp	2.5%	compass	3.3%
long sleeve	2.2%	graffiti	2.9%
snowdrift	2.0%	file folder	2.2%
dune	1.9%	**awning**	2.0%
art	1.9%	windshield wiper	2.0%
map	1.6%	circuitry	2.0%
compass	1.6%	art	2.0%
volleyball net	1.6%	shopping cart	1.6%
file folder	1.5%	ridge rope	1.5%
organdie	1.4%	dish rack	1.4%
herringbone pattern	1.3%	**goalpost**	1.4%
graffiti	1.3%	slide rule	1.4%
bookmark	1.2%	fluorescent lamp	1.3%
ocean	1.1%	map	1.2%
toothpick	1.0%	*automatic pistol*	1.2%

Most frequent categories chosen as the most confident by the Wolfram ImageIdentify Net for patches from parcel images. Percentages indicate how often chosen. Bolded categories are in the parcel-implausible subset. The most frequently chosen gun category is italicized.

IV. THREAT DETECTION

Before presenting our methods and results for unsupervised AD, we first assess how well our representations support supervised TD. This will determine an upper bound for AD performance: in simple terms whether the representations capture what is needed to distinguish the particular examples of anomaly that we test with (firearms) from SoC parcel contents. We note that this repeats previous work that has used the representations from object-in-photo CNNs for threat detection e.g. [14].

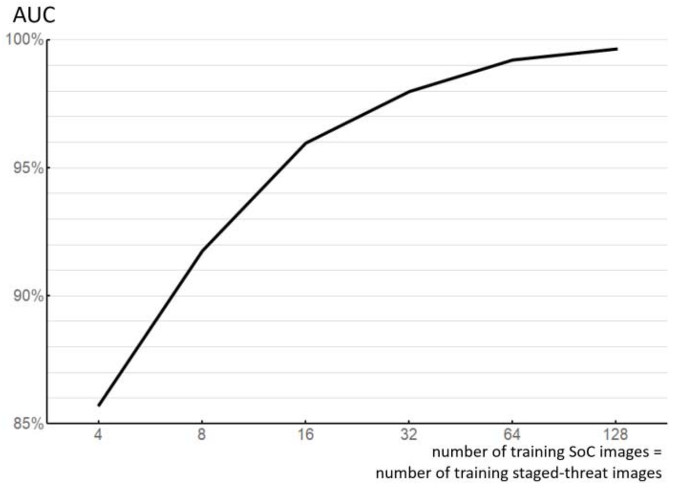


Fig. 6. Mean performance of supervised threat detection using the appearance representation, as a function of the training set size.

A. Appearance

To use the appearance representation (1024-D binary vectors per patch) for supervised TD, we trained a regularized logistic classifier using all patches from the same number of SoC images and threat images. For testing, the classifier was evaluated on a disjoint set of SoC and threat images. To produce a threat detection score for an image, the classifier was run for each patch separately, and the individual scores were averaged. Performance was quantified by AUC, which is the frequency with which a threat test image was given a higher score than a SoC test image. We prefer the AUC measure in this context as it makes no assumptions about what the rate of threats or anomalies will be, and does not require a detection threshold to be established. We report the mean AUC over repeated, random, train/test splits. Fig. 6 shows the mean performance as a function of the amount of training data. It shows that mean performance reached 99.6% (95% CI [99.5, 99.8]) when 128 SoC and 128 threat images were used for training. Higher performance looks likely if a larger training set was used, but we were limited by the 234 staged-threat images available. This confirms that the appearance representation is adequate for AD with the particular anomaly examples we are testing with.

B. Semantics

A supervised threat detection scheme based on the semantic representation (a 4315-D real-valued vector per image) can be constructed in the same manner as with the appearance representations. This has a performance of 99.5% when trained on 128 each of SoC and threat images, almost the same as the 99.6% for appearance. The slight shortfall can be attributed to the semantic layer being tuned to the appearance of firearms in photographs rather than X-ray images.

A different approach to using the semantic representation for TD of firearms is to exploit its known alignment to the threats that need to be detected: as listed in Table I, twelve of the dimensions are aligned to categories of gun.

Fig. 7 shows the means and standard deviations of the activations for different dimensions of the semantic representation for SoC and threat images. The figure shows that the

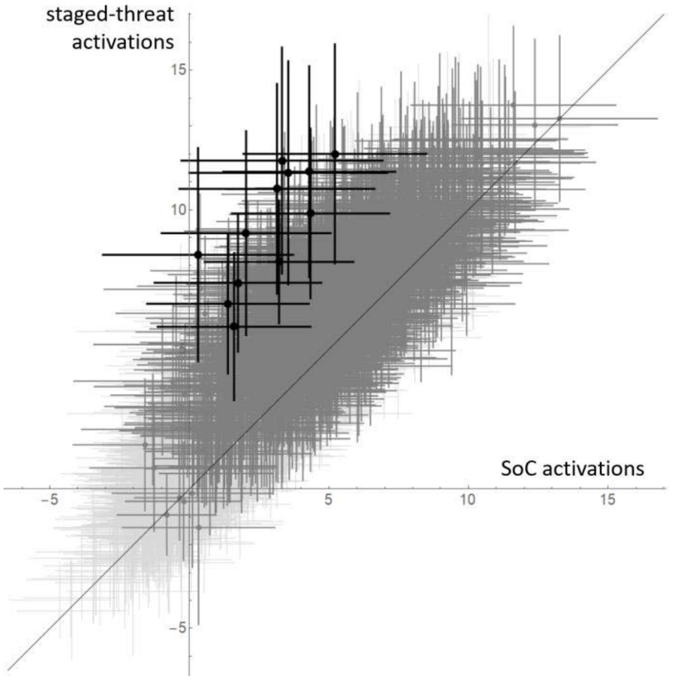


Fig. 7. Compares logit layer activations in SoC and threat images. A point marker for each of the 4315 dimensions marks the mean value within the datasets, its bars show the standard deviations. All 4315 dimensions are shown with light grey markers; on top of these, darker grey markers show for the parcel-plausible dimensions; on top of these, black markers show for gun categories.

representation values for threat images tend to have higher values than for SoC images (i.e. above the diagonal line) for all dimensions. For non-gun dimensions, this is possibly due to a tendency for the threat images to be slightly larger and busier. For gun dimensions, the elevation in values for threat images compared to SoC is clear, due to the presence of firearms in these images. The figure makes clear the challenge of semantic anomaly detection – while the presence of the anomalous categories is clear in a bulk comparison of threats to SoC images, per image these higher activations have to be detected in amongst the fluctuations of all the other categories.

A simple way to turn the elevated response to gun categories into a TD scheme is to score each image according to its maximum activity across the 12 gun categories. This gives an AUC of 93.7%, quite respectable for an out-of-the-box approach that has not been trained on, or fine-tuned for, X-ray images.

V. Anomaly Detection

Figure 8 gives an overview of the AD approaches we will describe in this section. The steps are: (i) representations are computed for patches of dual-view SoC images. (ii) for appearance anomalies the representations of patches within an image are kept separate, for semantic anomalies they are aggregated by a per-dimension maximum operation. (iii) a multinormal approximation of the distribution of SoC representations is constructed. (iv) the abnormality of a test image is assessed by computing the Mahalanobis distance of its representation(s) relative to that multinormal. (v) for appearance the most abnormal patch determines the abnormality of the image; for semantics the abnormality of the single maximum-aggregated representation is used.

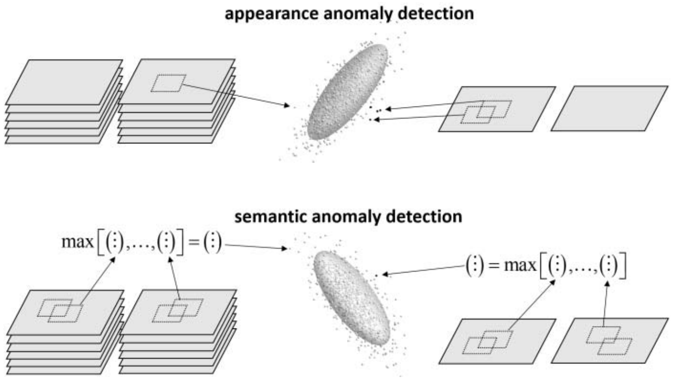


Fig. 8. Schematic overview of our approaches to Anomaly Detection. Left: training sets of dual-view SoC images. Centre: multinormal models of the distribution of SoC representations. Right: processing of a test image.

TABLE III
Performance of Baseline Features

feature	AUC
size	67.5%
attenuation	69.0%
busyness	72.3%

Before looking at our main AD approaches we establish baseline performance using simple engineered features (Table III). All perform above chance (50%). *Size*, measured in pixels, because the threat parcels tend to be larger than the SoC. *Attenuation*, computed by converting images to greyscale and summing the resulting values subtracted from the maximum value of 255, because the high density mass of the firearm tends to make threat images darker than SoC. *Busyness*, the sum of greyscale squared deviations from the mean, because inclusion of the firearm adds variation. Although all are above chance, the best performing achieves only 72.8%, showing that AD is not possible in this problem through simple approaches.

In the remainder of this section, in A we describe our approach for detection of appearance anomalies and give results when using 4096 SoC images as the normal class and the remaining SoC images and all threats images for testing; then the same for semantic anomalies (section B), and then for a combination of the two schemes to test whether they detect different things (C). We then present results on how performance of these schemes varies with the size of the normal data set (D), and the sensitivity of performance to the value of three hyper-parameters used in the schemes (E). In section F we report the effect of varying the pre-trained CNN used. In G we describe the processing times for this approach. In H we assess the use of a Generative Adversarial Network as an alternative to our principal way of modelling the distribution of normal data.

A. Appearance

Recall that the appearance representation is a 1024-D binary vector for each image patch. We model the distribution of

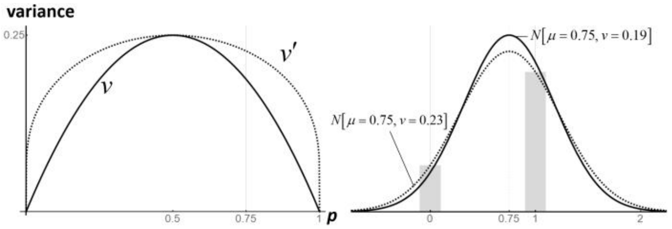


Fig. 9. Compares the naïve (solid) and adjusted (dotted) Gaussians used to model distributions of binary values. Left: variance of the modelling Gaussian as a function of the sample mean. Right: the modelling Gaussians (curves) for a distribution (grey bars) of binary values with mean 0.75. The bars coincide with the adjusted Gaussian at variate values 0 and 1.

these for SoC data using a multivariate Gaussian. The center of this distribution is simply the mean of the appearance vectors for the training dataset. For the covariance of the distribution, a naïve approach would be to use the sample covariance, but this ignores that we are using a Gaussian to model the distribution of binary, rather than scalar, values. As noted in section III.A, naïve Gaussian approximations of binary variables do not give the correct ratio between the probabilities of a 0 and a 1.

If in the training samples the fraction of 1s in some dimension is p then the naïve Gaussian approximation will have mean $\mu = p$ and variance $\upsilon = p\,(1-p)$. The ratio of the likelihoods that it assigns to a 1 and to a 0 is $e^{\frac{p-{}^{1}\!/_{2}}{p\,(1-p)}}$ different from the ratio $p\,(1-p)^{-1}$ in the sample. If the variance is determined instead according to $\upsilon' = (p - {}^{1}\!/_{2}) \big/ \ln\left(p\,(1-p)^{-1}\right)$ then the correct ratio results. Fig. 9 (left) shows the naïve and adjusted variances as a function of p. It shows that larger variances are set by the adjusted formula when p is close to 0 or 1 – recall though that the binarization threshold we use avoids values very close to these extremes. Fig. 9 (right) shows the naïve and adjusted Gaussians when $p = 0.75$ (also marked at the left).

The adjusted variance allows us to compute the modeling variance per dimension, but this gives only the diagonal values of the full covariance matrix ($\mathbf{\Sigma}$). We compute the covariance matrix from the correlation matrix ($\mathbf{C}$) and the vector of adjusted variances (V') by $\mathbf{\Sigma} := diag\left(\sqrt{V'}\right) \mathbf{C}\, diag\left(\sqrt{V'}\right)$. This covariance matrix, along with the vector of dimension means ($\vec{\mu}$) specifies the multivariate Gaussian we use to model the distribution of SoC appearance vectors. When assessing the abnormality of a test datum with representation $\vec{r}$ we can avoid computing its likelihood relative to that Gaussian, which is numerically problematic because of its high dimension, by instead computing the *log*-likelihood and ignoring the constant term. Hence the score is $(\vec{r} - \vec{\mu})^{\mathrm{T}}\, \mathbf{\Sigma}^{-1}\, (\vec{r} - \vec{\mu})$ (i.e. a Mahalanobis distance) with larger values signaling anomalies. We compute the anomaly score for an image, as the maximum of the anomaly scores for its patches i.e. an image is as anomalous as its most anomalous patch.

We have evaluated this AD scheme by splitting off a random subset of 4096 SoC images as training (from our pool of 5000 images), computing a multivariate Gaussian model for the distribution of patch representations for this data, and evaluating the anomaly scores of the remaining SoC images and all 234 threat images. We compute an AUC value from these scores. We repeat the training and testing multiple times using different random splits of the SoC into train and test sets, until the uncertainty of the mean has a 95% confidence less than 1% wide; and report the mean performance.

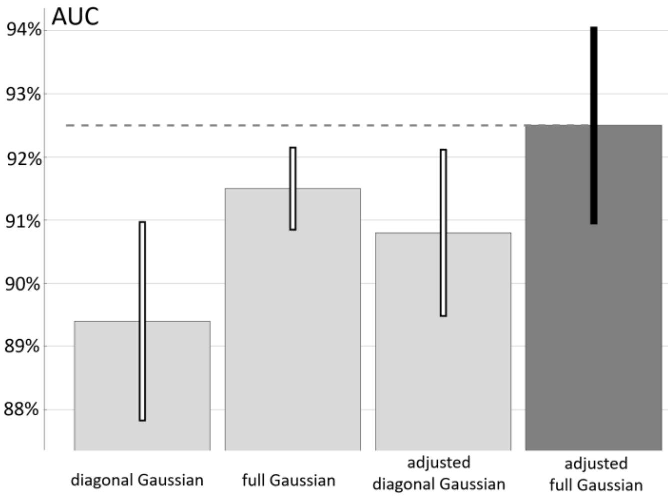


Fig. 10. Performance of different approaches to AD based on binarized appearance vectors. In all cases, the normal population was 4096 SoC parcels, and testing was on staged-threat parcels and held-out SoC parcels. Wide grey bars show mean performance using different random training and test sets with 95% confidence intervals (not shown) of less than ±1%. The best method ('adjusted full Gaussian') is shown with a darker bar and a 95% confidence interval showing the uncertainty of its *absolute* performance given the finite pool of training and test data. The other hollow error bars shows the 95% confidence interval of the *relative* performances of the sub-optimal methods compared to the best method; none cross the dashed line indicating that they are significantly worse.

To assess whether all aspects of our scheme deliver improved performance we also compute AUC scores using a diagonal rather than a full covariance for the Gaussian model, and without using the variance adjustment. The variant when we use the adjusted variance with a diagonal covariance is noteworthy. This is equivalent to a naïve Bayes scheme, where the likelihood of a datum is computed as the product of its likelihood in each dimension, computed as straightforward Bernoulli probabilities. This equivalence holds because the adjusted variance ensures that the probability of a 0 and 1 in each dimension are proportional to their rates in the training sample. We have confirmed that we do indeed get equal results if we compute the naïve Bayes scheme in a direct manner without use of Gaussian models.

Results are shown in Fig. 10: full covariance outperforms diagonal covariance, whether naïve or adjusted variance is used; adjusted variance outperforms naïve variance, whether full or diagonal covariance is used. The AUC for the main scheme – full covariance, adjusted variance – is 92.5%. The ±1.6% error bar shown in the figure (solid) is the 95% confidence of this estimate of performance given the finite size of the SoC and, particularly, the threat datasets. This was evaluated using multiple bootstrap re-samplings of the data, ensuring that multiple copies of an image were not split across train and test. The hollow error bars for the variant schemes show the uncertainty of their performance relative to the full scheme.

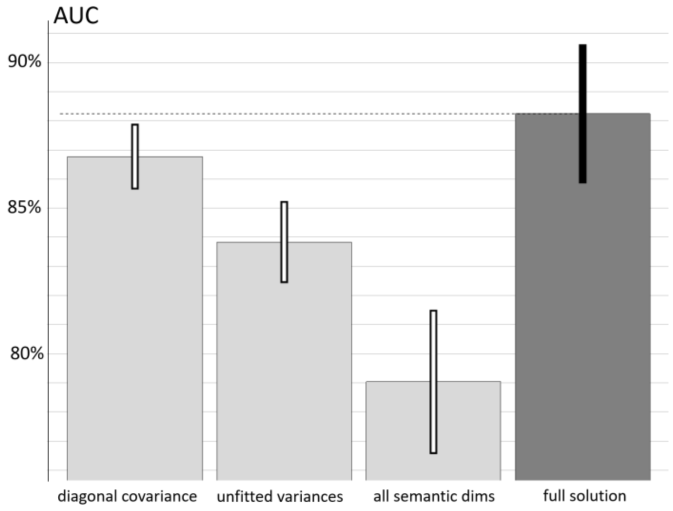


Fig. 11. Performance of different approaches to anomaly detection based on real-valued semantic vectors. Same details as Fig. 10.

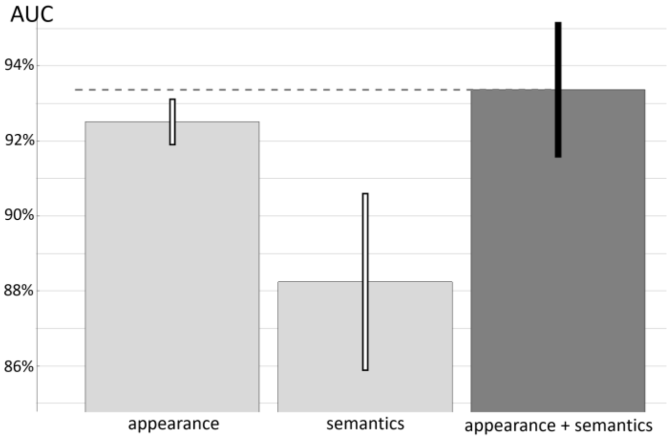


Fig. 12. Performance of combined vs. individual approaches to anomaly detection. Same details as Fig. 10.

Again this was evaluated using bootstrap resampling, with the difference between the performance of the variant and full schemes being computed for each re-sampling. None of these hollow error bars cross the performance level for the full scheme, demonstrating that their lower performance is significant at a 95% confidence level, rather than a quirk of the particular datasets we use.

B. Semantics

Recall that the appearance representation is a 4315-D real-valued vector for each image. As with appearance, we form a full covariance Gaussian model of the distribution of these. Since the vectors are real- rather than binary-valued there is no need to use a variance adjustment, but there are three non-standard changes we do make for this problem.

We consider only the subset of 996 parcel-plausible dimensions rather than the full 4315; because we can rule out as impossible anomalous appearance of these other categories (such as ‘giant redwood’).

We estimate the variance of each dimension from its sample mean and the fitted relationship shown in Fig. 4, rather than directly from the sample data.

For each test datum, when computing its likelihood relative to the model distribution, we consider only the n largest excursions away from the mean, rather than all excursions; where excursion is quantified by signed z-value. Smaller excursions in $\vec{\sigma}^{-1}(\vec{r}-\vec{\mu})$ are zeroed. We do this because we expect semantic anomalies to manifest as a positive increase in a small number of dimensions, tuned to categories similar to the anomalous object, rather than a diffuse pattern of increases and decreases across many dimensions. For a ball-park estimate of how many semantic dimensions might be co-activated for a typical semantic anomaly, we note that there are 12 categories of gun within the full 4315 (see Table I), so we set n to 12.

We evaluate detection of semantic anomalies using the protocol described for appearance. Results are shown in Fig. 11 for the full and variant schemes. The full scheme achieves an AUC performance of 88.2%, higher than the variants. Of the variants, using all semantic dimensions rather than the parcel-plausible subset has the greatest negative impact, while diagonal rather than full covariance has the least. The higher performance of the full scheme is statistically significant, as shown by none of the hollow error bars (showing the 95% confidence intervals of the variant scheme performance relative to the full scheme) crossing the dotted line. The absolute uncertainty of the full scheme is shown by the solid ±2.3% error bar which indicates its 95% confidence interval. All error bars were computed using bootstrap resampling with the same precautions used for appearance anomalies.

C. Combined

Although we have defined appearance and semantic anomalies differently, it is not self-evident that our approaches will detect different things. After all, a pattern of activations in the final pooling layer of the ImageIdentify net (which give rise to an appearance representation) can be considered a direction in appearance space, and the logit layer nodes (the activations of which are a semantic representation) also each correspond to a direction in that space. On the other hand, while the approach for detecting appearance anomalies requires all directions in this space to be equally monitored for unusual excursions from the mean; the approach for detecting semantic anomalies is provided with guidance on which directions correspond to semantically coherent categories, allowing it to closely monitor selected directions while ignoring others.

To test whether they do detect different things we combine the schemes, and see if the performance is improved. We combine the appearance anomaly score and the semantic anomaly score as a weighted sum. Since both scores are log probabilities, and so commensurate, we use an inverse-variance weighting scheme, which minimizes the variance of the weighted sum. The variances for this are determined from the spread of scores on the SoC training data. We determine them on the fly for each random set of training data, but they tend to around 0.92 for appearance and 0.08 for semantic i.e. a weight ratio of 11.5.

Results of the combined scheme, using 4096 SoC training images, are given in Fig. 12 along with comparison to the individual schemes. As with earlier results of this form, the error bars show 95% confidence intervals given the finite train and test datasets used; with the solid bar showing absolute performance, and the hollow bars showing performance relative to the combined scheme. Neither hollow error bar crosses the dashed line, so the combined performance of 93.4% is significantly higher than the individual performances of 92.5% for appearance and 88.2% for semantics. Thus we conclude that there is a small but significant non-overlap in what the appearance anomalies and semantic anomalies schemes detect.

The output of the combined scheme is illustrated in Fig. 13. The scatter plot shows appearance and semantic anomaly scores for the SoC data and the threat data, while the histograms show the combined score. For the combined score we have calculated the threshold value above which 90% of the staged threat scores lie, and displayed these in the two plots (green lines). This 90% detection scheme with a hard threshold gives a false alarm rate of 18% on the SoC data. The example images in the lower part of the figure show SoC images (top two rows) and staged-threat images (bottom two rows) with a range of combined scores. Which of these images have scores above and below the 90% detection threshold is indicated by the green polyline. Roughly reflecting the overall performance of the hard threshold, one of the ten threat images fails to register as an anomaly, and two of the ten SoC images do. It is noteable that the missed threat is less unusual looking than the others; and the false-alarm SoC images are more unusual looking.

D. Training Set Size

Our main results (Fig. 10-12) are computed with a training set of 4096 SoC images. This is large enough so that the covariance matrices used for the Gaussian models of the normal population AD are full rank, so invertible. With smaller training set sizes this is not the case. Using the pseudoinverse, rather than inverse, avoids this problem but causes performance to reduce sharply with decreasing training set size. Instead, we add a small multiple of the identity matrix to the sample covariance, to ensure invertiblity. With an appropriate weight, performance decreases smoothly and slowly with reduced training set size, and does not alter performance at the largest size.

Results of varying the size of the SoC training set are shown in fig 14. For all sizes, especially smaller ones, performance is averaged over multiple random splits of the SoC dataset into train and test portions.

The left plot shows the varying performance of appearance AD, and compares when full or diagonal covariance is used (in both cases using the adjusted variance calculation). It shows that diagonal covariance is superior for smaller training sets, and full covariance for larger. Additionally the performance for diagonal covariance plateaus earlier than for full covariance. All these observations are consistent with the greater number of parameters of the full covariance, supporting a more accurate model of the normal population density, but requiring extra data to reliably estimate.

The right plot compares appearance, semantics and combined anomalies. Appearance anomaly performance is seen to plateau in performance from 500 training images; while semantic anomaly performance plateaus from 2000 images. The combined appearance plus semantics scheme does not outperform appearance alone until 1000 images, and plateaus from 2000 images. We explain the failure of the combined scheme to outperform appearance at smaller training set sizes as due either to semantics not having anything to add to appearance until its performance is near maximum, or to a failure of the inverse-variance weighting scheme to identify effective score combination weights.

E. Hyper-Parameter Sensitivity

The anomaly detection methods have three hyper-parameters: the binarization threshold used to convert pooling layer activations into binary-valued appearance vector representations; the number of (largest) excursions from the mean considered in computing a semantic anomaly score; and the ratio between the weights used to combine appearance and semantic scores in the combined scheme. In supervised learning, the value of these would be set by tuning the performance score, using cross-validation to prevent over-fitting. In AD it is invalid to use anomalies ahead of test evaluation. Instead, in previous sections, we justified the particular values used. It is informative to compute the impact on performance had we used different values.

Fig. 15 shows the effect of varying the hyper-parameters. In all cases the full schemes (dark grey bars in Fig. 10-12) were trained on 4096 images. The hyper-parameter values used in the main results are indicated by vertical lines. The grey zone marks the range of values that give performance within 0.5% of the optimum, in all cases the used value lies within the grey zone. For the binarization threshold, the grey zone demarks a 6-fold range of threshold (i.e. 0.04 to 0.24); for the number of excursions from the mean a 5-fold range (i.e. 3 to 15); and for the combination weight ratio an 8-fold range (i.e. 8 to 64). So in all cases, the argued for values have been near optimum, and there is a useful latitude in the values of these hyper-parameters that achieve near peak performance.

F. Choice of CNN

The results presented were based on the Wolfram Image Identify (v11.1) CNN, chosen because of (i) its similarity to Inception V3 which is close to the state-of-the-art on ImageNet [56], and (ii) its large number of semantic categories. To assess the effect of this choice we have evaluated two alternative CNNs: ResNet-152 [57] and VGG-19 [58]. Data on these networks and their AD performance is presented in Table III. The ILSVRC'12 performance (2nd column) gives their rates at getting the correct answer as their top category, and within the top 5 on ImageNet – the figure given for the WolframNet is for Inception V3. The dimensions (3rd column) are of the final pooling layer for appearance and the parcel-plausible categories within their outputs for semantics.

The AD scores show that we made a good choice with the WolframNet. The pattern of results suggest that low dimensionality is desirable for appearance anomalies and high for semantic, but is not clear-cut.

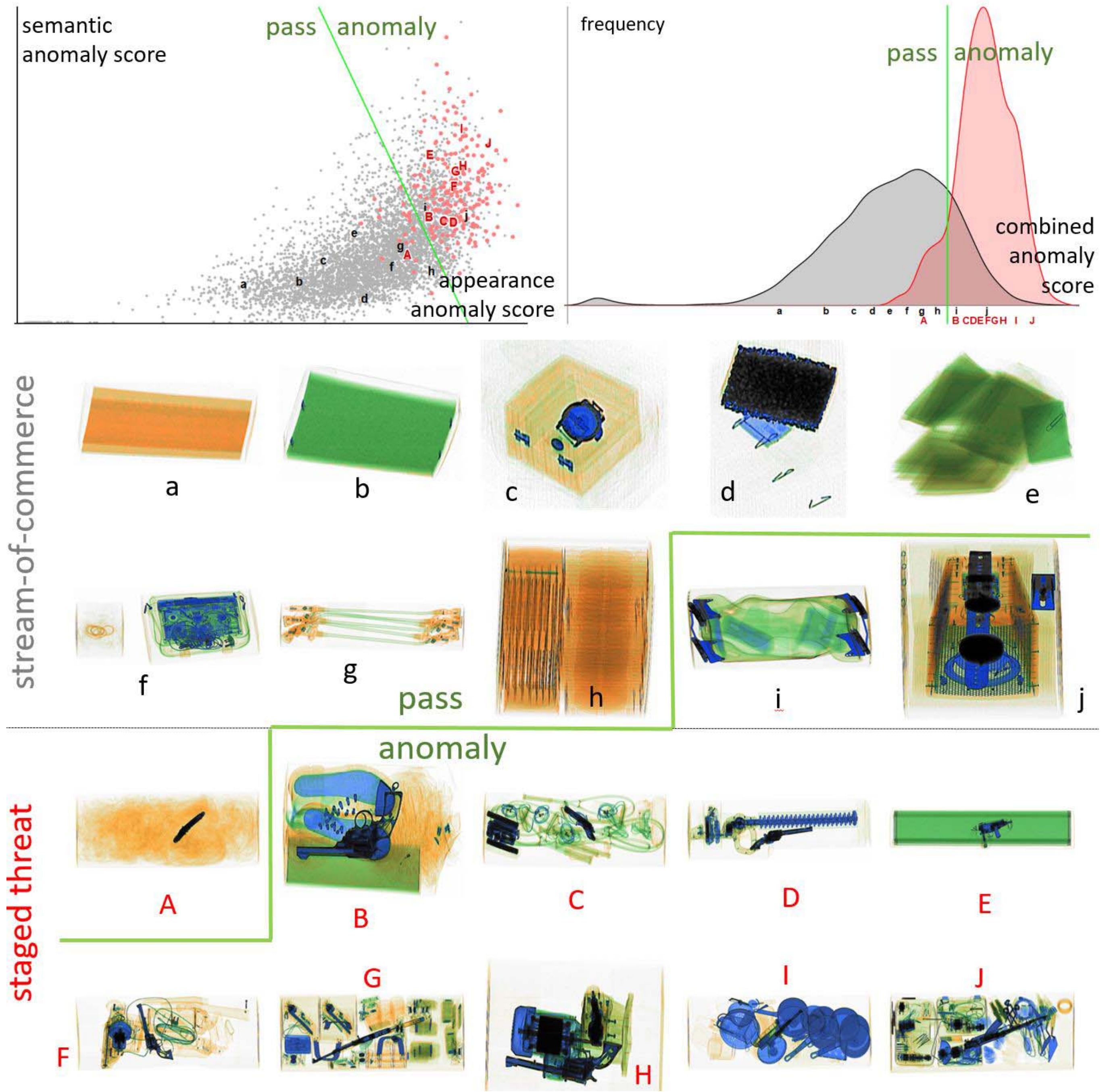


Fig. 13. Illustration of the combined scheme for anomaly detection. In all panels the green line marks the same hard threshold, images below this threshold 'pass', images above are considered 'anomalies'. The threshold is set to catch 90% of the staged-threat images, but triggers false alarms for 18% of SoC images. Top-left: appearance scores vs. semantic scores for SoC (grey) and threat (pink) images. Top-right: histograms of the combined scores, same colour scheme as left. Bottom four rows: example SoC and threat images. Letters labelling each image correspond to locations in the top row plots.

G. Computational Cost

In both training and testing the processing cost is dominated by the per-image computations, which can split into inferring the cropping boundary, dividing into patches, processing each patch by CNN, and thresholding (for appearance) or combining by max (for semantic) the extracted vectors, and computing Mahalanobis scores. Dual view images are decomposed, on average, into 26 patches. Assuming the CNN is pre-loaded, a Titan X GPU can run a patch through an Inception V3 CNN in 4ms. Total per image processing times of a second can be readily achieved, and with careful coding half a second should be possible on current hardware.

H. An AnoGAN Approach

Instead of modeling the distribution of SoC binary appearance vectors as a multinormal distribution, a plausible alternative is to use a Generative Adversarial Network (GAN) [59]. A GAN is pair of networks - a generator and a discriminator.

TABLE IV

CNN COMPARISON

CNN	ILSVRC '12 performance	Dim.	Appearance Anomalies	Semantic Anomalies
Wolfram Image Identify	~78.8% / ~94.4%	1024 / 996	92.5%	88.2%
ResNet-152	77.0% / 93.3%	2048 / 458	91.8%	79.7%
VGG-19	75.2% / 92.5%	4096 /458	84.6%	85.3%

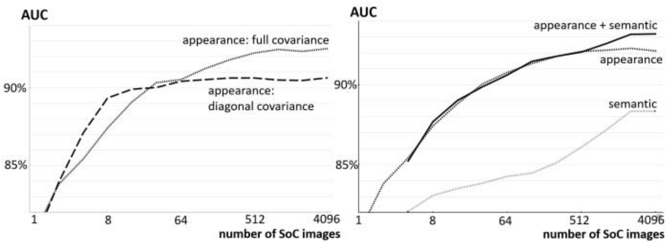


Fig. 14. Effect of training set size on performance. Left plot compares two schemes for appearance anomalies. Right plot compares individual and combined anomaly detection schemes.

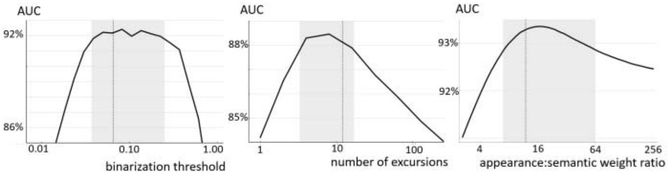


Fig. 15. Effect of hyper-parameters on performance. Left: appearance anomaly detection as a function of the binarization threshold. Middle: semantic anomaly detection as a function of the number of excursions from the mean. Right: combined anomaly performance as a function of the ratio of weights applied to the appearance and semantic scores.

The generator aims to synthesize data instances that appear to be drawn from the same population as a training set. The discriminator aims to distinguish between this synthetic data and real training data. During co-training of the networks, the generator is improved as it receives guidance from the discriminator on what aspects of the synthesized data need to be changed to make them more realistic, and the discriminator is improved as it is trained on more realistic synthetic data.

In [60] an AnoGAN method was presented for detecting anomalies in retinal OCT images. It is founded on the fact that the generator operates by transforming a latent variate (typically with a high-dimensional isotropic multinormal distribution) into synthetic data instances. Having trained a GAN on normal data, test data is processed by discovering the latent variate that best reproduces them when fed into the generator. The degree to which that discovered latent variate is outside of the normal range of the latent variates works as an anomaly score.

We adapted the AnoGAN method to detect appearance anomalies in parcels using, not raw image data, but the binary appearance vectors we have used for our main approach. Since these vectors do not have a spatial structure we used discriminator and generator networks with fully connected, rather than convolutional layers, as in the original GAN paper [59]. We experimented in the architecture of the networks using a validation subset of our data, settling on a generator network that receives a 200-D latent variate input that then passes through layers with 128, 256, 512 and 1024 units. The discriminator had 128, 256, 512, 1024 and 1 units in its layers. We followed [61] for choice of non-linear units, batch normalization strategy and training method.

Other than using per-patch anomaly scores from the GAN, our assessment of the approach replicated that used with our main multinormal approach (section V.A). The resulting AUC was 73.5%, higher than our baseline approaches (Table III) but far short of the multinormal approach (92.5%). This is a disappointing result given the good performance in [60]. We suspect that the difference is the difference in diversity of the normal class in the two problems. Possibly an AnoGAN can perform well for our problem with its very diverse normal class, and it does have the *potential* to capture dependencies higher-order than pairwise which the multinormal approach never can, but it will require further network architecture engineering to achieve this.

VI. SUMMARY & CONCLUSIONS

We presented approaches for detection of appearance and semantic anomalies in X-ray security images. Our approach is sophisticated in the representations it uses, and simple in how anomalies are detected given those representations.

For both types of anomaly we used representations extracted from layers of an object classification CNN trained on photographic images. The anomaly status of test images was assessed by computing the likelihoods of their representations relative to full covariance multivariate Gaussian models of the distribution of representations of normal data.

The schemes were assessed using parcel images. A stream-of-commerce dataset was taken as the normal class. A staged-threat dataset of parcels with normal contents plus a firearm were considered as example anomalies.

Anomaly detection performance increased with the size of the training set. For appearance anomalies this plateaued at 500 images, achieving an AUC score of 92.5%. For semantic anomalies, performance plateaued at 2000 training images, at which size it achieved an AUC of 88.2%. Combining the approaches yielded a slight improvement to an AUC of 93.4%, but as this was shown to be statistically significant it was confirmed that the two schemes were not detecting identical aspects of image structure. The 95% confidence interval of the performance of the combined scheme was ±2%, given the finite datasets available.

When implemented with a hard detection threshold, the best performing scheme was able to detect 90% of staged firearms as anomalies while raising false alarms on 18% of stream-of-commerce data. While this is much lower than can achieved by direct threat detection of firearms based on supervised training it is possibly good enough to find a role within screening, or as a supplement to a threat detection system (possibly operating with a higher threshold) able to pick-up anomalies that do not correspond to a specified list of threat items.

The scheme for detecting Appearance Anomalies has a limited capacity for localizing the detections as illustrated in fig 16. This shows an anomaly score for each pixel computed as the average of the scores for the windows containing the pixel. Since the windows are large the localization is crude

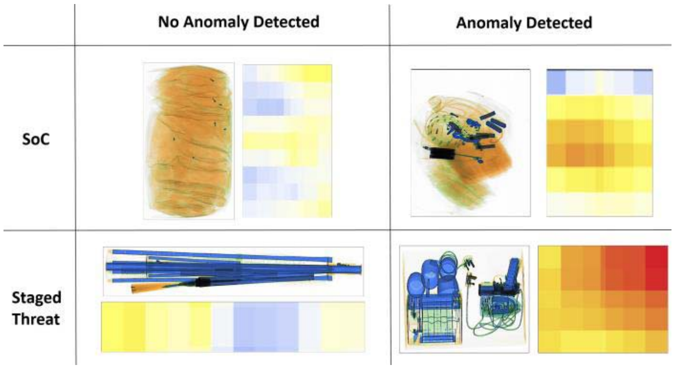


Fig. 16. Example localization maps for anomaly detection, paired with a view from a dual view image. Colors toward red indicate high anomaly scores, towards blue low. From left-to-right and top-to-bottom the examples show a true negative, a false alarm a failed detection and a true positive.

but would have some use to an operator. A similar output for semantic anomalies has not been attempted as the max operation across the vectors for each patch, which is performed before computing the anomaly score, makes this difficult.

There are five potential avenues for improving performance of the system. We consider these in order of their use in the algorithms.

1. *Different CNN architecture*. New architectures, with improved object detection performance are regularly being proposed [54], [62]; a better performing network may have generally more effective features for appearance anomalies, as well as better tuning for semantic anomalies.
2. *X-ray trained CNN*. The currently used CNN was trained on photographic images and, as can be seen from Table II, has only weak object classification performance on X-ray images. A CNN trained to do diverse semantic classification of X-ray data would be expected to produce much more effective semantic representations, but may in addition produce better appearance representations as the aspects of image structure that support semantic classification in X-rays and photographs are likely to have some differences.
3. *Different distribution modeling*. If the normal population distribution is not exactly a multivariate Gaussian, then modeling it as such will misrepresent aspects of it. For the binary appearance representations, if there are dimensional dependencies beyond pairwise the multivariate Gaussian model will not represent them. For the real-valued semantic representations, the Gaussian form of the dimensional marginals does not guarantee that the population distribution is Gaussian. A model-free approach to modeling the distribution, for example kernel-density-estimation, might be able to perform better.
4. *Outlier detection*. Instead of a density modelling approach, a boundary-based method of outlier detection could be effective. For example 1-SVM for semantics, or isolation trees for appearance.
5. *Increased training data*. In the current system, performance has plateaued before the maximum size of training set with which we have worked. However it cannot be ruled out that if any of 1-4 above yield improved performance, then the system might be able to extract value from a larger training dataset.

Of these options, we rate 3 and 4 as likely to yield only small benefit, 1 as moderate benefit, and 2 as high benefit; 5 is plausible when any improvement has been made. Although 2 is by far the most promising route to improvement, the difficulties in this approach need to be appreciated. Assembling massive labeled datasets for training photo classifiers is facilitated by the ubiquity of cameras and the internet, and has many economic drivers. A viable route to creating a comparable labeled dataset for X-ray images is not obvious. Potentially this blockage can be side-stepped by learning a 'translation' between photo and x-ray appearance with a smaller 'parallel text' dataset.

We conclude that anomaly detection in X-ray security images can achieve a useful level of performance by utilizing the representational power of photo-appearance object classification networks; and that there is good potential to achieve much better performance using an X-ray trained network, but the data sourcing challenges of this are considerable.

References

[1] E. Taylor, "Supermarket self-checkouts and retail theft: The curious case of the SWIPERS," *Criminol. Criminal Justice*, vol. 16, pp. 552–567, Apr. 2016.

[2] G. Zentai, "X-ray imaging for homeland security," *Int. J. Signal Imag. Syst. Eng.*, vol. 3, no. 1, pp. 13–20, 2010.

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

[4] A. Schwaninger, D. Hardmeler, and F. Hofer, "Aviation security screeners visual abilities & visual knowledge measurement," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 20, no. 6, pp. 29–35, Jun. 2005.

[5] H. Vogel and D. Haller, "Luggage and shipped goods," *Eur. J. Radiol.*, vol. 63, no. 2, pp. 242–253, 2007.

[6] C. C. von Bastian, A. Schwaninger, and S. Michel, "Do multi-view X-ray systems improve X-ray image interpretation in airport security screening?" *Zeitschrift Arbeitswissenschaft*, vol. 62, pp. 165–173, Jan. 2008.

[7] R. D. R. Macdonald, "Design and implementation of a dual-energy X-ray imaging system for organic material detection in an airport security application," *Proc. SPIE*, vol. 4301, pp. 31–42, Apr. 2001.

[8] J. R. Pierce and P. K. Schott, "Concording U.S. harmonized system categories over time," Nat. Bur. Econ. Res., Tech. Rep. w14837, 2009.

[9] K. E. Yoo and Y. C. Choi, "Analytic hierarchy process approach for identifying relative importance of factors to improve passenger security checks at airports," *J. Air Transp. Manage.*, vol. 12, pp. 135–142, May 2006.

[10] T. W. Rogers, N. Jaccard, E. J. Morton, and L. D. Griffin, "Automated X-ray image analysis for cargo security: Critical review and future promise," *J. X-ray Sci. Technol.*, vol. 25, no. 1, pp. 33–56, 2017.

[11] N. Jaccard, T. W. Rogers, E. J. Morton, and L. D. Griffin, "Detection of concealed cars in complex cargo X-ray imagery using deep learning," *J. X-ray Sci. Technol.*, vol. 25, pp. 323–339, 2017.

[12] M. Caldwell, M. Ransley, T. W. Rogers, and L. D. Griffin, "Transferring X-ray based automated threat detection between scanners with different energies and resolution," *Proc. SPIE*, vol. 10441, p. 104410F, Oct. 2017.

[13] M. Roomi and R. Rajashankarii, "Detection of concealed weapons in X-ray images using fuzzy K-NN," *Int. J. Comput. Sci., Eng. Inf. Technol.*, vol. 2, no. 2, pp. 187–196, 2012.

[14] D. Mery, E. Svec, M. Arias, V. Riffo, J. M. Saavedra, and S. Banerjee, "Modern computer vision techniques for X-ray testing in baggage inspection," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 4, pp. 682–692, Apr. 2017.

[15] 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 Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2203–2215, Sep. 2018.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[17] J. Zhang *et al.*, "Joint shape and texture based X-ray cargo image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2014, pp. 266–273.

[18] J. Tuszynski, J. T. Briggs, and J. Kaufhold, "A method for automatic manifest verification of container cargo using radiography images," *J. Transp. Secur.*, vol. 6, no. 4, pp. 339–356, 2013.

[19] F. Rembold, C. Atzberger, I. Savin, and O. Rojas, "Using low resolution satellite imagery for yield prediction and yield anomaly detection," *Remote Sens.*, vol. 5, no. 4, pp. 1704–1733, 2013.

[20] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 25, no. 7, pp. 5–28, Jul. 2010.

[21] Y. Tarabalka, T. V. Haavardsholm, I. Kåsen, and T. Skauli, "Real-time anomaly detection in hyperspectral images using multivariate normal mixture models and GPU processing," *J. Real-Time Image Process.*, vol. 4, no. 3, pp. 287–300, 2009.

[22] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1975–1981.

[23] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1446–1453.

[24] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *J. Imag.*, vol. 4, no. 2, p. 36, 2018.

[25] T. W. Rogers, N. Jaccard, E. J. Morton, and L. D. Griffin, "Detection of cargo container loads from X-ray images," in *Proc. 2nd IET Int. Conf. Intell. Signal Process.*, 2015, pp. 1–6.

[26] J. T. A. Andrews, T. Tanay, E. J. Morton, and L. D. Griffin, "Transfer representation-learning for anomaly detection," in *Proc. ICML*, 2016, pp. 1–5.

[27] J. T. A. Andrews, E. J. Morton, and L. D. Griffin, "Detecting anomalous data using auto-encoders," *Int. J. Mach. Learn. Comput.*, vol. 6, no. 1, pp. 21–26, 2016.

[28] Y. Zheng and A. Elmaghraby, "A vehicle threat detection system using correlation analysis and synthesized X-ray images," *Proc. SPIE*, vol. 8709, p. 87090V, Jun. 2013.

[29] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. 15, 2009.

[30] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Comput. Netw.*, vol. 51, no. 12, pp. 3448–3470, Aug. 2007.

[31] European Commission, "Concealment methods," in *Good Practice Guide for Sea Container Control*. 2002.

[32] L. Greenemeier, "Exposing the weakest link: As airline passenger security tightens, bombers target cargo holds," *Sci. Amer.*, Nov. 2010. [Online]. Available: https://www.scientificamerican.com/article/aircraft-cargo-bomb-security/

[33] J. Barrett and T. Westbrook, "Islamic State behind Australians' foiled Etihad meat-mincer bomb plot—Police," in *Reuters*. 2017. [Online]. Available: https://uk.reuters.com/article/uk-australia-security-raids-idUKKBN1AJ35Z

[34] B. G. Brogdon, H. Vogel, and J. D. McDowell, *A Radiologic Atlas of Abuse, Torture, Terrorism, and Inflicted Trauma*. Boca Raton, FL, USA: CRC Press, 2003.

[35] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Int. Conf. Artif. Neural Netw.*, 2011, pp. 52–59.

[36] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture IE*, vol. 2, pp. 1–18, Dec. 2015.

[37] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[38] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[39] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.

[40] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2008, pp. 413–422.

[41] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection," in *Applications of Data Mining in Computer Security*. Boston, MA, USA: Springer, 2002, pp. 77–101.

[42] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Rec.*, 2000, pp. 93–104.

[43] R. Laxhammar, G. Falkman, and E. Sviestins, "Anomaly detection in sea traffic—A comparison of the Gaussian mixture model and the kernel density estimator," in *Proc. 12th Int. Conf. Inf. Fusion (FUSION)*, 2009, pp. 756–763.

[44] L. J. Latecki, A. Lazarevic, and D. Pokrajac, "Outlier detection with kernel density functions," in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit.*, 2007, pp. 61–75.

[45] B. A. Turlach, "Bandwidth selection in kernel density estimation: A review," in *CORE and Institut de Statistique*. 1993. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.6770

[46] N. Ben Amor, S. Benferhat, and Z. Elouedi, "Naive Bayes vs decision trees in intrusion detection systems," in *Proc. ACM Symp. Appl. Comput.*, 2004, pp. 420–424.

[47] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *Proc. Eur. Conf. Mach. Learn.*, 1998, pp. 4–15.

[48] L. Jiang, D. Wang, Z. Cai, and X. Yan, "Survey of improving Naive Bayes for classification," in *Proc. Int. Conf. Adv. Data Mining Appl.*, 2007, pp. 134–145.

[49] D. R. Cox and N. Wermuth, "A note on the quadratic exponential binary distribution," *Biometrika*, vol. 81, no. 2, pp. 403–408, 1994.

[50] Y. Freund and D. Haussler, "Unsupervised learning of distributions on binary vectors using two layer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 1992, pp. 912–919.

[51] D. R. Cox and E. J. Snell, *Analysis of Binary Data*, vol. 32. Boca Raton, FL, USA: CRC Press, 1989.

[52] D. R. Cox, "The analysis of multivariate binary data," *Appl. Statist.*, vol. 21, no. 2, pp. 113–120, 1972.

[53] *Wolfram ImageIdentify Net V1*. Accessed: Feb. 21, 2017. [Online]. Available: https://resources.wolframcloud.com/NeuralNetRepository/resources/Wolfram-ImageIdentify-Net-V1

[54] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, 2017, p. 12.

[55] C. Olah *et al.*, "The building blocks of interpretability," *Distill*, vol. 3, no. 3, p. e10, 2018.

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

[58] K. Simonyan and A. Zisserman. (Sep. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[59] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

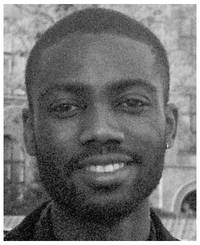
[60] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2017, pp. 146–157.

[61] A. Radford, L. Metz, and S. Chintala. (Nov. 2015). "Unsupervised representation learning with deep convolutional generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1511.06434

[62] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3859–3869.

Lewis D. Griffin received the B.A. degree in mathematics and philosophy from Oxford University, U.K., in 1988, and the Ph.D. degree from the University of London in 1995. His Ph.D. thesis was on "descriptions of image structure" in the area of computational vision. Following positions at Aston University (Vision Sciences) and Kings College London (Imaging Sciences), he has been with University College London in computer science since 2005, where is currently a Reader. His research interests include image structure, color vision, machine learning, and biomedical modeling, with applications in security science, biomedicine, and geoscience.

Matthew Caldwell worked in commercial software development for 15 years in games and finance before returning to academia to undertake a Ph.D. in biophysics. His doctoral research combined computational modeling and experimental electrophysiology to study synaptic communication in the brain. He subsequently did post-doctoral work on the models of human cerebral physiology before joining Lewis Griffin's research group in 2017 to apply machine learning techniques to problems in security imaging.



Jerone T. A. Andrews received the M.Sc. degree (Hons.) in mathematics from King's College London in 2013, the M.Res. degree in security science and Ph.D. degree in computer science from University College London (UCL) in 2014. He is currently a Research Associate with the Department of Computer Science, UCL, and a Holder of the Royal Academy of Engineering U.K. Intelligence Community Research Fellowship from 2018 to 2020. His Ph.D. thesis was on Representation Learning for Anomaly Detection in Computer Vision. His research interests include unsupervised representation learning, transfer learning, and anomaly detection.

Helene Bohler received the B.S. degree in mathematics from Bergen University in 2017, and the M.S. degree in machine learning from University College London in 2018. She is currently with Spacemaker as a Data Scientist, where she develops advanced optimization tools for urban site development. Her main fields of interest are reinforcement learning and deep learning for machine vision applications.